

طرق التحليل الإحصائي

متعدد المتغيرات

إعداد

الأستاذ الدكتور زياد رشاد الراوي

نشر بدعم من المعهد العربي للتدريب والبحوث الإحصائية

2017

الطبعة الأولى

2017م

المملكة الأردنية الهاشمية
رقم الإيداع لدى دائرة المكتبة الوطنية
(2017/10/5553)

❖ يتحمل المؤلف كامل المسؤولية القانونية عن محتوى مصنفه ولا يعبّر هذا المصنف عن رأي دائرة المكتبة الوطنية أو أي جهة حكومية أخرى.
❖ تم إعداد بيانات الفهرسة والتصنيف الأولية من قبل دائرة المكتبة الوطنية

فهرس المحتويات

تقديم

إن التزايد المتسارع والملاحظ في الحاجة لإستخدام الطرق الإحصائية بمستوياتها وجوانبها المختلفة لتشمل الظواهر الحياتية منها والعلمية يصاحبه تطور وتنوع في أبعاد هذه الطرق الإحصائية لضمان شمول جميع هذه الظواهر وبما يناسب نوعية وحجم البيانات المتاحة من جهة، ومستويات وأبعاد التحليل الإحصائي المراد إنجازه من جهة أخرى.

وإنطلاقاً من رسالة المعهد العربي للتدريب والبحوث الإحصائية في العمل المتواصل على تطوير قدرات العاملين في مجال التحليل الإحصائي لأنواع مختلفة من البيانات، يقوم المعهد بتنظيم دورات تدريبية ويشجع على إنجاز دراسات وأبحاث في هذا الإتجاه وفي اتجاهات إحصائية أخرى. وفي هذا الإطار يأتي دعمه لإصدار هذا الكتاب من تأليف الأستاذ الدكتور زياد رشاد الراوي تحت عنوان " طرق التحليل الإحصائي متعدد المتغيرات ".

إن هذه الطرق من التحليل الإحصائي تعتبر القمة في تحليل البيانات لأنها تأخذ في الإعتبار التنوع في البيانات والتعددٍ في المتغيرات التي تمثلها. ولكونها يغلب عليها التعامل مع المصفوفات فقد ارتأى المؤلف تخصيص فصل موسع حول هذا الموضوع في بداية الكتاب. من ناحية أخرى، ونظراً لتعدد المتغيرات التي تتضمنها طرق التحليل هذه فقد يواجه الباحث حالة تعدد أنماط القياس لهذه المتغيرات، ولتدارك الأمر في هذا الجانب يتضمن الكتاب أيضاً موضوع القياس بمختلف أنواعه (الإسمي، الترتيبي، الفئوي، النسبي) مع توضيح ذلك بالنسبة لجميع المتغيرات التي تتضمنها كل طريقة من طرق التحليل المتعدد والتي تناولها هذا الكتاب .

لا يسعني إلا أن أتقدم بالشكر الجزيل للأستاذ الدكتور زياد رشاد الراوي على هذه المساهمة العلمية المتميزة، وآمل أن يكون المعهد قد قدم بذلك مرجعاً هاماً في التحليل الإحصائي متعدد المتغيرات لمختلف الإحصائيين والمهتمين بذات الموضوع.

والله ولي التوفيق

عبد العزيز معلمي
المدير العام

مدخل عام General Introduction

إن المقصود بمتعدد المتغيرات هو التعامل مع حالة وجود أكثر من متغير واحد سواءً فيما يتعلق بالمتغيرات (العوامل) التوضيحية (وتسمى بالمستقلة أحياناً Independent) من جهة، أو متغيرات الإستجابة (وتسمى بالمعتمِدة Dependent) من جهة أخرى.

والبيانات متعددة المتغيرات تظهر في جميع تفرعات العلوم تقريباً. وفي الغالب نجد أن جميع البيانات التي يتم جمعها من خلال الوحدة التجريبية Experimental Unit وتحليلها من قبل الباحثين يمكن تصنيفها بكونها بيانات متعددة المتغيرات. والمقصود بالوحدة التجريبية هنا هي أي حالة أو عنصر يمكن قياسه أو تقييمه بطريقة ما. والبيانات متعددة المتغيرات تظهر هنا متى ما قام الباحث بقياس أو تقييم أكثر من خاصية أو سمة واحدة لكل وحدة تجريبية. وهذه الخواص أو السمات تسمى عادة بالمتغيرات من قبل الإحصائيين.

وطرق التحليل متعدد المتغيرات في غاية الأهمية لكونها تساعد الباحثين في تكوين إستنتاج فيما يخص مجاميع كبيرة ومتداخلة ومعقدة أحياناً من البيانات تتضمن عدداً كبيراً من المتغيرات مأخوذة من عدد كبير من الوحدات التجريبية. إن ضرورة وفائدة إستخدام طرق التحليل متعدد المتغيرات تزداد بشكلٍ طردي مع زيادة عدد الوحدات التجريبية أو عدد المتغيرات المأخوذة عنها للتحليل.

وغالباً ما يكون الهدف من إستخدام التحليل متعدد المتغيرات هو تلخيص الكمية الكبيرة من البيانات من خلال عدد صغير نسبياً من المعلمات Parameters. وبالتالي فإن الوظيفة الرئيسية لغالبية الأساليب متعددة المتغيرات هي التبسيط.

من جانب آخر، فإن التحليل متعدد المتغيرات غالباً ما يرتبط مع إيجاد علاقات ما بين:

1) متغيرات الإستجابة Response Variables
2) الوحدات التجريبية Experimental Units
3) كل من متغيرات الإستجابة والوحدات التجريبية

إن غالبية أساليب التحليل متعدد المتغيرات تميل إلى كونها ذات طبيعة إستكشافية لحالةٍ ما بدلاً من كونها تأكيدية لتلك الحالة. وهذا يعني ميلها إلى خلق الفرضيات الإحصائية بدلاً من اختبارها.

ولتوضيح هذه النقطة، إفترض حالة كون باحثٍ ما لديه (50) خمسون متغيراً مقاسة على أكثر من (2000) ألفي وحدة تجريبية. إن الطرق الإحصائية الإعتيادية تتطلب من الباحث البدء بإدراج بعض الفرضيات أولاً ومن ثم قيامه بجمع البيانات، ومن بعد ذلك إستخدام هذه البيانات لتثبيت الميل الإحتمالي لتبني (قبول) هذه الفرضيات أو نفي ثبوت صحتها (رفضها).

والحالة البديلة التي غالباً ما تظهر هنا هي حالة إمتلاك الباحث لكمية كبيرة من البيانات ويساوره الشعور فيما إذا كانت هنالك ثمة معلومات ذات أهمية ضمن هذه البيانات. إن أساليب التحليل متعدد المتغيرات غالباً ما تكون مفيدة في عملية إستكشاف ضمن البيانات لمحاولة الوصول إلى نوع من القناعة بأن ثمة معلومات ذات قيمة ومفيدة تنطوي عليها مجموعة البيانات هذه.

تصنيف طرق التحليل متعدد المتغيرات

أحد الفروقات الأساسية ما بين الطرق متعددة المتغيرات يكمن في تصنيفها إلى صنفين رئيسيين من حيث أساليب التحكم في التحليل وهما:

1) أساليب تحكّم المتغيرات Variable – directed techniques

وتشمل تلك التي تتعامل بشكلٍ رئيسي مع العلاقات التي من الممكن ظهورها ضمن متغيرات الإستجابة Response Variables التي يتم قياسها. ومثال ذلك:

- التحليلات المعتمدة على مصفوفات معامل الإرتباط Correlation Matrices
- تحليل المركبات الرئيسية Principle Components Analysis
- التحليل العاملي Factor Analysis
- تحليل الإرتباط القويم (إرتباط المجاميع) Canonical Correlation Analysis

2) أساليب التحكم الشخصي Individual – directed techniques

وتشمل تلك التي تتعامل بشكلٍ رئيسي مع العلاقات التي من الممكن ظهورها ضمن الوحدات التجريبية و/أو الأشخاص الخاضعين للقياس. ومثال ذلك:

- التحليل المميز Discriminant Analysis
- التحليل العنقودي Cluster Analysis
- تحليل التباين المتعدد Multivariate Analysis of Variance (MANOVA)

تكوين متغيرات جديدة

قد نجد في كثير من الأحيان أنه من المفيد تكوين متغيرات جديدة لكل وحدة تجريبية ليكون بالإمكان عمل مقارنة فيما بينها بطريقة أكثر سهولة. هذه المتغيرات الجديدة عبارة عن دوال تتضمن جميع المتغيرات الأصلية المعتمدة في التجربة. إن العديد من الطرق متعددة المتغيرات تساعد الباحث في تكوين متغيرات جديدة تتسم بمزايا مرغوبة. مثل هذه الطرق هي المركبات الرئيسية، التحليل العاملي، تحليل الإرتباط القويم، التحليل المميز القويم.

عندما نبدأ التحليل

ما أن يبدأ الباحث التفكير في إجراء تحليل لمجموعة جديدة من البيانات، فإن أسئلة عديدة حول هذه البيانات يجب أن تؤخذ بالإعتبار. ومن هذه الأسئلة المهمة:

1) هل هنالك أية جوانب في البيانات يمكن إعتبارها غريبة أو غير إعتيادية؟
2) هل يمكن الإفتراض بأن البيانات تتوزع طبيعياً Normality؟
3) هل هنالك أي جانب منها خارج الطبيعي Abnormality ؟
4) هل هنالك أية قيم شاردة (شاذة) Outliers في البيانات؟ نقصد هنا بالقيمة الشاذة لوحدة التجربة حيث المتغيرات المقاسة منها تبدو وكأنها غير متناسقة مع القياسات المأخوذة لوحدات أخرى.

وفي هذا الكتاب سنتناول المواضيع التالية:

- تحليل الإنحدار المتعدد
- المركبات الرئيسية
- الإنحدار اللوجستي
- تحليل التباين المتعدد
- التحليل المميز
- التحليل العنقودي
- الإرتباط القويم
- التحليل العاملي

ولضرورة فهم التعامل مع المصفوفات والعمليات المرتبطة بها، نرى من الضروري البدء في تغطية هذا الموضوع قبل البدء في المواضيع الرئيسية أعلاه.

وبما أن المتغيرات التي سوف نتعامل معها تتبع مستويات مختلفة من القياس، لذلك سيتم تناول هذا الموضوع أيضاً بعد استعراض فصل المصفوفات.

عمليات المصفوفات
Matrix Operations

يتم تعريف المصفوفة بأبعادها التي تشير إلى عدد الصفوف وعدد الأعمدة ضمن هذه المصفوفة. فالمصفوفة $A_{m,n}$ ترمز للمصفوفة A ذات أبعاد m من الصفوف و n من الأعمدة. أي أنها تتضمن mxn من العناصر a_{ij} وتكتب بالصيغة التالية:

$$A_{m,n} = \begin{bmatrix} a_{11} & a_{12} & .... & a_{1n} \\ a_{21} & a_{22} & .... & a_{2n} \\ \\ a_{m1} & a_{m2} & .... & a_{mn} \end{bmatrix}$$

ويمكن تدوير هذه المصفوفة بإستبدال الصفوف مع الأعمدة وتسمى بالمصفوفة المدورة Transpose of matrix ونرمز لها بالرمز $A'_{m,n}$ وتكون بالصيغة التالية:

$$A'_{m,n} = \begin{bmatrix} a_{11} & a_{21} & .... & a_{m1} \\ a_{12} & a_{22} & .... & a_{m2} \\ \\ a_{1n} & a_{2n} & .... & a_{mn} \end{bmatrix}$$

مثال: إذا كانت لدينا المصفوفة:

$$A_{3,2} = \begin{bmatrix} 2 & 3 \\ 0 & -1 \\ 1 & 0 \end{bmatrix} \Rightarrow A'_{2,3} = \begin{bmatrix} 2 & 0 & 1 \\ 3 & -1 & 0 \end{bmatrix}$$

عملية الجمع والطرح في المصفوفات

لنفترض المصفوفتين $A_{m,n}$ و $B_{m,n}$ حيث أن:

$$A_{m,n} = \begin{bmatrix} a_{11} & a_{12} & .... & a_{1n} \\ a_{21} & a_{22} & .... & a_{2n} \\ \\ a_{m1} & a_{m2} & .... & a_{mn} \end{bmatrix} , \quad B_{m,n} = \begin{bmatrix} b_{11} & b_{12} & .... & b_{1n} \\ b_{21} & b_{22} & .... & b_{2n} \\ \\ b_{m1} & b_{m2} & .... & b_{mn} \end{bmatrix}$$

فإن ناتج جمع المصفوفتين أو طرحهما يتم بجمع أو طرح ما بين كل عنصرين متقابلين (a_{ij} و b_{ij}) من المصفوفتين لتكوين مصفوفة جديدة $C_{m,n}$ بعناصر c_{ij} وعلى النحو التالي:

$$A_{m,n} + B_{m,n} = \begin{bmatrix} (a_{11}+b_{11}) & (a_{12}+b_{12}) & .... & (a_{1n}+b_{1n}) \\ (a_{21}+b_{21}) & (a_{22}+b_{22}) & .... & (a_{2n}+b_{2n}) \\ & & & \\ (a_{m1}+b_{m1}) & (a_{m2}+b_{m2}) & .... & (a_{mn}+b_{mn}) \end{bmatrix}$$

أما عملية الطرح فتتم باستبدال إشارة الجمع بإشارة الطرح. أي أن:

$$A_{m,n} - B_{m,n} = \begin{bmatrix} (a_{11}-b_{11}) & (a_{12}-b_{12}) & .... & (a_{1n}-b_{1n}) \\ (a_{21}-b_{21}) & (a_{22}-b_{22}) & .... & (a_{2n}-b_{2n}) \\ & & & \\ (a_{m1}-b_{m1}) & (a_{m2}-b_{m2}) & .... & (a_{mn}-b_{mn}) \end{bmatrix}$$

مثال: لنفترض المصفوفتين:

$$\mathbf{A_{3,3}} = \begin{bmatrix} 1 & -1 & 3 \\ 0 & 4 & -2 \\ 1 & 0 & 5 \end{bmatrix} \quad , \quad \mathbf{B_{3,3}} = \begin{bmatrix} 0 & 1 & 2 \\ 2 & 0 & 3 \\ 1 & 1 & 4 \end{bmatrix}$$

$$A_{3,3} + B_{3,3} = \begin{bmatrix} 1 & 0 & 5 \\ 2 & 4 & 1 \\ 2 & 1 & 9 \end{bmatrix}$$

$$A_{3,3} - B_{3,3} = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -5 \\ 0 & -1 & 1 \end{bmatrix}$$

ملاحظة:

عند عملية الجمع أو الطرح، يشترط أن تكون المصفوفتان بنفس الأبعاد من حيث عدد الصفوف وعدد الأعمدة.

عملية الضرب في المصفوفات

هنا يشترط أن يكون عدد صفوف إحدى المصفوفتين مساوياً لعدد أعمدة الأخرى. وتتم العملية بضرب عناصر العمود (j) من المصفوفة الثانية بعناصر الصف (i) من المصفوفة الأولى وبشكل متقابل ترتيبياً وجمع حاصل الضرب ليكون العنصر (ij) من المصفوفة الناتجة وبالشكل التالي:

$$A_{m,n} \times B_{n,p} = \begin{bmatrix} a_{11} & a_{12} & .... & a_{1n} \\ a_{21} & a_{22} & .... & a_{2n} \\ \\ a_{m1} & a_{m2} & .... & a_{mn} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} & .... & b_{1p} \\ b_{21} & b_{22} & .... & b_{2p} \\ \\ b_{n1} & b_{n2} & .... & b_{np} \end{bmatrix}$$

$$= \begin{bmatrix} \sum a_{1i}b_{i1} & \sum a_{1i}b_{i2} & .... & \sum a_{1i}b_{ip} \\ \sum a_{2i}b_{i1} & \sum a_{2i}b_{i2} & .... & \sum a_{2i}b_{ip} \\ \\ \sum a_{mi}b_{i1} & \sum a_{mi}b_{i2} & .... & \sum a_{mi}b_{ip} \end{bmatrix}$$

مثال: لنفترض المصفوفتين:

$$A = \begin{bmatrix} 2 & 1 & 3 \\ 0 & -1 & 2 \\ 3 & 0 & 4 \\ 5 & 1 & -1 \end{bmatrix} \quad , \quad B = \begin{bmatrix} -2 & 1 \\ 0 & 5 \\ 3 & 2 \end{bmatrix}$$

$$A_{4,3} \times B_{3,2} = \begin{bmatrix} 2 & 1 & 3 \\ 0 & -1 & 2 \\ 3 & 0 & 4 \\ 5 & 1 & -1 \end{bmatrix} \times \begin{bmatrix} -2 & 1 \\ 0 & 5 \\ 3 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} (-4+0+9) & (2+5+6) \\ (0+0+6) & (0-5+4) \\ (-6+0+12) & (3+0+8) \\ (-10+0-3) & (5+5-2) \end{bmatrix} = \begin{bmatrix} 5 & 13 \\ 6 & -1 \\ 6 & 11 \\ -13 & 8 \end{bmatrix}$$

حالة ضرب المصفوفة بالقيمة الثابتة c

في هذه الحالة ينتج لدينا مصفوفة بعناصر جديدة عبارة عن العناصر الأصلية مضروبة بـ القيمة الثابتة C.

$$c \times A_{m,n} = c\begin{bmatrix} a_{11} & a_{12} & .... & a_{1n} \\ a_{21} & a_{22} & .... & a_{2n} \\ \\ a_{m1} & a_{m2} & .... & a_{mn} \end{bmatrix} = \begin{bmatrix} ca_{11} & ca_{12} & .... & ca_{1n} \\ ca_{21} & ca_{22} & .... & ca_{2n} \\ \\ ca_{m1} & ca_{m2} & .... & ca_{mn} \end{bmatrix}$$

مثال: لنفترض المصفوفة:

$$A = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 4 & 1 \\ -1 & 2 & -3 \end{bmatrix}$$

$$5A = \begin{bmatrix} 15 & 5 & 10 \\ 0 & 20 & 5 \\ -5 & 10 & -15 \end{bmatrix}$$

محددة المصفوفة |A| Determinant

لنفترض المصفوفة المربعة $A_{n,n}$ حيث $n = 3$ على سبيل المثال ولغرض التبسيط:

$$A_{3,3} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

حيث أن المحددة |A| تكون عند إستخدامنا الصف الأول كعامل ضرب:

$$|A| = a11M11 - a_{12}M_{12} + a_{13}M_{13}$$

$$= a11C11 + a_{12}C_{12} + a_{13}C_{13}$$

حيث $C_{ij} = (-1)^{i+j} M_{ij}$

وأن:

$$M_{11} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$$

بعد حذف الصف الأول والعمود الأول، وبشكلٍ عام، فإن:

$$M_{ij} = |A^{-}|$$

أي أنها تساوي محددة ما يتبقى من المصفوفة A بعد حذف الصف (i) والعمود (j) (أي محددة ما سميناه A^{-}).

ملاحظة :

في أعلاه، إستخدمنا الصف الأول كعامل ضرب وبإستطاعتنا إستخدام أي صف أو عمود لنفس الغرض ونختار الأسهل في الحسابات. وبصورة عامة، يمكن كتابة قيمة محددة المصفوفة $A_{n,n}$ على الشكل التالي:

1- بإستخدام الصف (i) كعامل ضرب تكون:

$$\det(A) = |A| = \sum_{j=1}^{n} a_{ij} C_{ij}$$

2- بإستخدام العمود (j) كعامل ضرب تكون:

$$\det(A) = |A| = \sum_{i=1}^{n} a_{ij} C_{ij}$$

مثال: إفترض المصفوفة:

$$A_{3,3} = \begin{bmatrix} 3 & 4 & -1 \\ 2 & -4 & 5 \\ 0 & 1 & -6 \end{bmatrix}$$

وبإستخدام العمود الأول تكون:

$$|A| = 3 \begin{vmatrix} -4 & 5 \\ 1 & -6 \end{vmatrix} - 2 \begin{vmatrix} 4 & -1 \\ 1 & -6 \end{vmatrix} + 0 \begin{vmatrix} 4 & -1 \\ -4 & 5 \end{vmatrix}$$

$= 3[(-4)(-6) – (5)(1)] – 2[(4)(-6) – (-1)(1)] + 0.0$

$= 3(19) - 2(-23)$

$= 103$

ولو إستخدمنا الصف الأخير، يكون لدينا:

$$|A| = 0 \begin{vmatrix} 4 & -1 \\ -4 & 5 \end{vmatrix} - 1 \begin{vmatrix} 3 & -1 \\ 2 & 5 \end{vmatrix} - 6 \begin{vmatrix} 3 & 4 \\ 2 & -4 \end{vmatrix}$$

$= 0.0 – 1[(3)(5) – (-1)(2)] – 6[(3)(-4) – (4)(2)]$

$= 0.0 – 17 + 120$

$= 103$

ملاحظة:

لأجل التبسيط في عملية تحديد قيمة المحددة لأي مصفوفة مربعة، من المستحسن إختيار الصف أو العمود الذي يتضمن أصفاراً أكثر. وعادة ما يطلق على هذه الطريقة بالإختيار الذكي للصف أو العمود.

مثال: لو كانت لدينا المصفوفة التالية:

هنا من المستحسن إختيار العمود الثاني لكثرة الأصفار فيه. أي أنه سيكون لدينا:

$$|A| = 0 + 1\begin{vmatrix} 1 & 0 & -1 \\ 1 & -2 & 1 \\ 2 & 0 & 1 \end{vmatrix} + 0 + 0$$

وهنا يتبقى لدينا فقط الحد الثاني وسنختار العمود الثاني أيضاً لنفس السبب ليكون لدينا:

$$|A| = 0 + 1(-2)\begin{vmatrix} 1 & -1 \\ 2 & 1 \end{vmatrix} + 0$$

$$= -2\,[1 + 2] = -6$$

ملاحظة:

في حالة عدم وجود أصفار كثيرة في أحد الصفوف أو الأعمدة لتسهيل العملية الحسابية في إستخراج محددة المصفوفة، لا يزال لدينا خياراً آخر لتسهيل حصول ذلك وهو تطبيق طريقة "لابلاس Laplace" وكما هو مبين في أدناه.

طريقة لابلاس Laplace

تستخدم هذه الطريقة لإحداث أصفارفي صفٍ ما أو عمودٍ ما للمصفوفة في حالة عدم وجودها وذلك قبل حساب محددة المصفوفة، وعن طريق ضرب ذلك الصف بقيمة ثابتة وجمعها مـع صف آخر أو إجراء نفس الشئ مع ذلك العمود وتكون هذه العملية مجزية خاصة في حالة القيم الكبيرة نسبياً لعناصر المصفوفة. وقد تتوالى هذه العملية لأكثر من مرة واحدة وحسب الضرورة والمثال التالي يوضح ذلك.

مثال: لو كانت لدينا المصفوفة A

$$A = \begin{bmatrix} 4 & 3 & 0 & -1 \\ 9 & 7 & 2 & -3 \\ 4 & 0 & 2 & -1 \\ 3 & -1 & 4 & 5 \end{bmatrix}$$

$$|A| = \begin{vmatrix} 4 & 3 & 0 & -1 \\ 9 & 7 & 2 & -3 \\ 4 & 0 & 2 & -1 \\ 3 & -1 & 4 & 5 \end{vmatrix}$$

وبضرب العمود الأخير بالعدد (4) وجمعه إلى العمود الأول يكون هذا مساوياً إلى:

$$= \begin{vmatrix} 0 & 3 & 0 & -1 \\ -3 & 7 & 2 & -3 \\ 0 & 0 & 2 & -1 \\ 23 & -1 & 4 & 5 \end{vmatrix}$$

والآن نقوم بضرب العمود الأخير بالعدد (3) وجمعه إلى العمود الثاني فيكون هذا مساوياً إلى:

$$= \begin{vmatrix} 0 & 0 & 0 & -1 \\ -3 & -2 & 2 & -3 \\ 0 & -3 & 2 & -1 \\ 23 & 14 & 4 & 5 \end{vmatrix}$$

$$= 1 \begin{vmatrix} -3 & -2 & 2 \\ 0 & -3 & 2 \\ 23 & 14 & 4 \end{vmatrix}$$

لأننا إستخدمنا الصف الأول

ويمكننا إستخراج العدد (2) مضروباً بالعمود الأخير (وهذا يصح في حالة المحددة) ليكون لدينا:

$$= 2 \begin{vmatrix} -3 & -2 & 1 \\ 0 & -3 & 1 \\ 23 & 14 & 2 \end{vmatrix}$$

وبضرب العمود الأخير بالعدد (3) وجمعه للعمود الثاني يصبح لدينا:

$$= 2 \begin{vmatrix} -3 & 1 & 1 \\ 0 & 0 & 1 \\ 23 & 20 & 2 \end{vmatrix}$$

$$= 2\,(-1) \begin{vmatrix} -3 & 1 \\ 23 & 20 \end{vmatrix} = -2\,[(-3)(20) - (1)(23)] = 166$$

ملاحظة:

ليس من الضروري الإبقاء على التعامل مع الأعمدة وإنما يمكن التغيير نحو الصفوف أيضاً بشكلٍ مماثل وفي أي خطوة.

بعض النظريات المفيدة في موضوع محددة المصفوفة

1) إذا كان كل عنصر ضمن صف أو عمود مضروباً في عدد ثابت k، فإن محددة المصفوفة تكون قيمتها مضروبة في k.
2) قيمة محددة المصفوفة تساوي صفراً في حالة:
   أ- كل عناصر أي صف أو أي عمود أصفاراً.
   ب- كان صفان أو عمودان متشابهين.
   ت- وجود علاقة نسبية بين أي صفين أو عمودين.
3) في حالة تغيير موقع عمودين أو صفين مع بعضهما، فإن قيمة المحددة تتغير إشارتها فقط.
4) لا تتغير قيمة المحددة إذا:
   أ- تم كتابة الأعمدة صفوفاً أو الصفوف أعمدة.
   ب- أضفنا لكل عنصر في صفٍ ما العنصر المقابل له في صفٍ آخر مضروباً في العدد k. (نفس الشئ ينطبق بالنسبة للأعمدة).

ملاحظة:

إن وجود محددة للمصفوفة من عدمها تعتمد على وجود معكوس للمصفوفة من عدمها وكما سنرى ذلك لاحقاً.

حالة المصفوفة المثلثية Triangular Matrix

تعرف المصفوفة المثلثية بأنها تلك التي تكون جميع عناصرها ضمن المثلث فوق المحور Diagonal أو الذي تحته أصفاراً. أي أنها مثل:

$$A = \begin{bmatrix} a_{11} & 0 & .. & .. & 0 \\ a_{21} & a_{22} & 0 & .. & 0 \\ & & & & 0 \\ a_{n1} & a_{n2} & .. & ... & a_{nn} \end{bmatrix}$$

وفي مثل هذه الحالة، فإن قيمة محددة المصفوفة A عبارة عن حاصل ضرب جميع العناصر المحورية (a_{ii}) فقط. أي أن:

$$|A| = \prod_{i=1}^{n} a_{ii} = a_{11}a_{22}....a_{nn}$$

مثال: لنفترض المصفوفة التالية:

$$A = \begin{bmatrix} 3 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -2 & 4 \end{bmatrix}$$

$|A| = (3)(1)(4) = 12$

معكوس المصفوفة Matrix Inverse

لنفترض أن للمصفوفة A معكوس هو A^{-1} حيث أن:

$A^{-1} = adj(A)/|A|$

ولذلك فإن معكوس المصفوفة يتحدد بوجود محددة لها. وحيث أن:

$adj(A) = C'$

علماً أن:

$$C = \begin{bmatrix} c_{11} & c_{12} & .... & c_{1n} \\ c_{21} & c_{22} & .... & c_{2n} \\ & & & \\ c_{n1} & c_{n2} & .... & c_{nn} \end{bmatrix}$$

وأن: $C_{ij} = (-1)^{i+j} M_{ij}$

وأن M_{ij} هي محددة المصفوفة A بعد حذف الصف (i) والعمود (j) منها ومثلما سبق وذكرنا ذلك في بداية الفصل.

مثال: لنفترض المصفوفة A

$$A = \begin{bmatrix} 3 & 2 & -1 \\ 1 & 6 & 3 \\ 2 & -4 & 0 \end{bmatrix}$$

$M_{11} = \begin{vmatrix} 6 & 3 \\ -4 & 0 \end{vmatrix} = 12 \qquad C_{11} = (-1)^{1+1} M_{11} = 12$

$M_{12} = \begin{vmatrix} 1 & 3 \\ 2 & 0 \end{vmatrix} = -6 \qquad C_{12} = 6$

$$M_{13} = \begin{vmatrix} 1 & 6 \\ 2 & -4 \end{vmatrix} = -16 \qquad C_{13} = -16$$

$$M_{21} = \begin{vmatrix} 2 & -1 \\ -4 & 0 \end{vmatrix} = -4 \qquad C_{21} = 4$$

$$M_{22} = \begin{vmatrix} 3 & 3-1 \\ 2 & 0 \end{vmatrix} = 2 \qquad C_{22} = 2$$

$$M_{23} = \begin{vmatrix} 3 & 2 \\ 2 & -4 \end{vmatrix} = -16 \qquad C_{23} = 16$$

$$M_{31} = \begin{vmatrix} 2 & -1 \\ 6 & 3 \end{vmatrix} = 12 \qquad C_{31} = 12$$

$$M_{32} = \begin{vmatrix} 3 & -1 \\ 1 & 3 \end{vmatrix} = 10 \qquad C_{32} = -10$$

$$M_{33} = \begin{vmatrix} 3 & 2 \\ 1 & 6 \end{vmatrix} = 16 \qquad C_{33} = 16$$

وبذلك فإن:

$$C = \begin{bmatrix} 12 & 6 & -16 \\ 4 & 2 & 16 \\ 12 & -10 & 16 \end{bmatrix}$$

$$adj\,(A) = C' = \begin{bmatrix} 12 & 4 & 12 \\ 6 & 2 & -10 \\ -16 & 16 & 16 \end{bmatrix}$$

وحيث أن:

$$|A| = 3\begin{vmatrix} 6 & 3 \\ -4 & 0 \end{vmatrix} - 2\begin{vmatrix} 1 & 3 \\ 2 & 0 \end{vmatrix} + (-1)\begin{vmatrix} 1 & 6 \\ 2 & -4 \end{vmatrix}$$

$$= 3(12) - 2(-6) - 1(-16)$$

$$= 36 + 12 + 16 = 64$$

وبذلك تكون معكوسة المصفوفة A كما يلي:

$$A^{-1} = adj(A)/\,|A|\, = (1/64)\begin{bmatrix} 12 & 4 & 12 \\ 6 & 2 & -10 \\ -16 & 16 & 16 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{12}{64} & \frac{4}{64} & \frac{12}{64} \\ \frac{6}{64} & \frac{2}{64} & \frac{-10}{64} \\ \frac{-16}{64} & \frac{16}{64} & \frac{16}{64} \end{bmatrix}$$

ملاحظة:

إن جدوى إستخراج معكوس المصفوفة تظهر في غالبية عمليات التحليل متعدد المتغيرات والتي يمكن تمثيل كل أو جزء من بياناتها بمصفوفة. ونحن نعلم بأن جميع طرق التحليل متعدد المتغيرات تبرز فيها مثل هذه الحالة.

كما أن هنالك طرقاً مثل تحليل الإنحدار المتعدد وتحليل الإنحدار اللوجستي المتعدد تبرز فيهما حالة نظام المعادلات الخطية والتي يمكن إستخدام A^{-1} فيها إضافة إلى إمكانية اتباع طرق أخرى لحل مثل هذه المعادلات وإستخراج قيم المجاهيل فيها. وسوف نتناول ذلك من خلال ما يلي:

حل نظام معادلات خطية Solving system of linear equations

لنفترض أنه لدينا المعادلتين الخطيتين:

$2X - Y = 5$

$X + 2Y = -5$

وهذه يمكن تمثيلها بالمصفوفات وكما يلي:

$$A\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 5 \\ -5 \end{bmatrix}$$

وبالتالي فإننا نريد معرفة قيم X و Y اللتان تحققان صحة المعادلات هذه. وهذا يعني أن:

$$\begin{bmatrix} x \\ y \end{bmatrix} = A^{-1}\begin{bmatrix} 5 \\ -5 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}^{-1}\begin{bmatrix} 5 \\ -5 \end{bmatrix}$$

$$= \frac{1}{5}\begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 5 \\ -5 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 5 \\ -15 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

طريقة غاوس – جوردن Gouss - Jordan للمعادلات الخطية

ويمكن تطبيق هذه الطريقة على المثال أعلاه بالشكل التالي:

1- نضع الصيغة التالية:

$$\left[\begin{array}{cc|c} 2 & -1 & 5 \\ 1 & 2 & -5 \end{array}\right]$$

2- نبدأ بإجراء التحويلات على الجزء الأيسر من الصيغة هذه بحيث نصل بها للشكل التالي:

$$\left[\begin{array}{cc|c} 1 & 0 & c_1 \\ 0 & 1 & c_2 \end{array}\right] = \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & -3 \end{array}\right]$$

أي أننا نجري التحويلات المناسبة على المصفوفة A والتي تمثل الطرف الأيسر حتى تحويلها إلى مصفوفة الوحدة (مصفوفة بمحور جميع عناصره "1" مع بقية العناصر خارجه أصفاراً). هذه التحويلات بالطبع تأخذ مداها على الجانب الأيمن تلقائياً. والنتيجة هي أن:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

وفيما يلي الإسلوب المتبع لإجراء هذه العملية:

نبدأ بنفس الصيغة:

$$\left[\begin{array}{cc|c} 2 & -1 & 5 \\ 1 & 2 & -5 \end{array}\right]$$

ولأننا بحاجة إلى الرقم "1" للعنصر a_{11}، فإنه من المناسب إستبدال الصفين أحدهما مكان الآخر. بهذه العملية لم يتم تغيير أي شئ سوى أننا إعتبرنا المعادلة الثانية هي الأولى. أي أنه لدينا:

$$\left[\begin{array}{cc|c} 1 & 2 & -5 \\ 2 & -1 & 5 \end{array}\right]$$

وبضرب الصف الأول بالمقدار (-2) وجمعه للصف الثاني نحصل على:

$$\left[\begin{array}{cc|c} 1 & 2 & -5 \\ 0 & -5 & 15 \end{array}\right]$$

وبضرب الصف الثاني بالمقدار (1/5-) يكون لدينا:

$$\left[\begin{array}{cc|c} 1 & 2 & -5 \\ 0 & 1 & -3 \end{array}\right]$$

وبضرب الصف الثاني بالمقدار (2-) وجمعه للصف الأول يكون لدينا:

$$\left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & -3 \end{array}\right]$$

وبما أننا توصلنا لما نريده فإن الحل هو:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

مثال آخر: لنفترض أن مجموعة المعادلات الخطية التالية:

$X + Y = 5$

$-2X - Y + 2Z = -10$

$3X + 6Y + 7Z = 14$

هذه المعادلات تعطينا الصيغة التالية:

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & 5 \\ -2 & -1 & 2 & -10 \\ 3 & 6 & 7 & 14 \end{array}\right]$$

وبضرب الصف الأول بالمقدار (2) وجمعه للصف الثاني، وكذلك بضرب الصف الأول بالمقدار (3-) وجمعه للصف الثالث نحصل على:

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & 5 \\ 0 & 1 & 2 & 0 \\ 0 & 3 & 7 & -1 \end{array}\right]$$

وبضرب الصف الثاني بالمقدار (1-) وجمعه للصف الأول، وكذلك بضرب الصف الثاني بالمقدار (2-) وجمعه للصف الثالث نحصل على:

$$\left[\begin{array}{ccc|c} 1 & 0 & -2 & 5 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & -1 \end{array}\right]$$

وبضرب الصف الثالث بالمقدار (2) وجمعه للصف الأول، وكذلك بضرب الصف الثالث بالمقدار (2-) وجمعه للصف الثاني نحصل على:

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1 \end{array}\right]$$

وبما أننا توصلنا لما نريده فإن الحل هو:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ -1 \end{bmatrix}$$

ملاحظة:

عندما يكون لدينا حلاً لنظام المعادلات، فإن ذلك يعني وجود معكوسة للمصفوفة A وهذا بدوره يعني حتمية وجود محددة لهذه المصفوفة. وفي حالات أخرى غير ذلك، فقد لا نجد حلاً لنظام المعادلات هذا. ومثال ذلك في الحالتين التاليتين:

(1) عند وجود حالة الترابط الخطي ما بين أية معادلتين وهذه الحالة تدعى dependent system ومثال ذلك لو كانت لدينا الصيغة:

$$\left[\begin{array}{cc|c} 2 & -4 & 4 \\ 1 & -2 & 2 \end{array}\right]$$

فإنه بضرب الصف الأول بالمقدار (1/2-) وجمعه للصف الثاني، سيصبح لدينا:

$$\left[\begin{array}{cc|c} 2 & -4 & 4 \\ 0 & 0 & 0 \end{array}\right]$$

وهذه عبارة عن معادلة واحدة بمتغيرين مجهولين وهنا يصعب إيجاد حلاً لها بالنسبة لقيم المتغيرين.

(2) عند غياب حالة التناسق ما بين المعادلات مما يعني عدم صحة معادلة أو أكثر عند نفس القيم للمجاهيل. وهذه الحالة تدعى inconsistent system. ومثال ذلك لو كانت لدينا الصيغة التالية:

$$\left[\begin{array}{cc|c} 2 & -5 & 10 \\ 2/5 & -1 & 7 \end{array}\right]$$

فإنه بضرب الصف الأول بالمقدار (1/5-) وجمعه للصف الثاني، سيصبح لدينا:

$$\left[\begin{array}{cc|c} 2 & -5 & 10 \\ 0 & 0 & 5 \end{array}\right]$$

وهذا يعني أن 5 = 0.0 وهو غير منطقي وبالتالي لا يوجد حلاً لهذا النظام.

طريقة الحذف Elimination Method لحل نظام المعادلات الخطية

وهذه الطريقة تعتمد إسلوب حل كل معادلتين آنيتين بطريقة الحذف التوافقى للمتغيرات والإنتهاء بمتغير واحد نجد قيمته ومن ثم الرجوع عكسياً لتعويض هذه القيمة في معادلة ذات متغيرين تتضمن متغيراً واحداً إلى جانب هذا المتغير ليتم تحديد قيمة المتغير الثاني. وتتكرر هذه العملية على معادلة أخرى تتضمن متغيراً ثالثاً إلى جانب هذين المتغيرين لغرض إيجاد قيمته. وتستمر هذه العملية تباعاً حتى الإنتهاء من تحديد قيم جميع المتغيرات في مجموعة المعادلات هذه.

وفيما يلي مثالاً لتوضيح مجريات هذه الطريقة:

مثال: لنفترض نظام المعادلات الخطية التالية:

A $\quad 2X + y - 3Z = -7$

B $\quad 3X - 2Y + Z = 11$

C $\quad -2X - 3Y - 2Z = 3$

ملاحظة:

لقد تم تسمية المعادلات هذه بالحروف A و B و C لغرض تسهيل الرجوع اليها وسنستمر بتسمية المعادلات الجديدة بنفس الإسلوب.

ومن أجل إيجاد الحل، سنتبع الخطوات التالية بالنسبة لهذه المجموعة:

1) حذف المتغير Y من المعادلتين A و B بعد ضرب المعادلة A بالمقدار (2) والذي سيتم التنويه بذلك إزاء المعادلة نفسها. أي أننا سنحذف المتغير Y من المعادلتين 2A و B بطريقة الجمع أو الطرح لتصبح لدينا المعادلة الجديدة D :

(2A) $\quad 4X + 2y - 6Z = -14$

B $\quad 3X - 2Y + Z = 11$

بالجمع

D 7 X - 5Z = -3

2) حذف المتغير Y من المعادلتين A و C بنفس الإسلوب:

(3A) 6X + 3y – 9Z = - 21

C - 2X – 3Y - 2Z = 3

بالجمع

E 4A - 11Z = - 18

3) حذف المتغير X من المعادلتين D و E بنفس الإسلوب:

(-4D) -28 X + 20Z = 12

(7E) 28X - 77Z = - 126

بالجمع

- 57Z = - 114

ومنها نجد أن:Z = 2

4) وبالتعويض عن هذه القيمة في المعادلة D نحصل على X = 1.
5) وبتعويض هاتين القيمتين بالمعادلة A نحصل على Y = -3.
6) وبذلك تكون نتيجة الحل النهائي لهذا النظام هي:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}$$

قاعدة كريمر Cramer's rule لحل نظام المعادلات الخطية

لنفترض المعادلات الخطية التالية بصيغة المصفوفات:

$$\begin{bmatrix} 1 & 0 & 2 \\ -3 & 4 & 6 \\ -1 & -2 & 3 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 30 \\ 8 \end{bmatrix}$$

وهذا يعني أن المصفوفة A هي:

$$A = \begin{bmatrix} 1 & 0 & 2 \\ -3 & 4 & 6 \\ -1 & -2 & 3 \end{bmatrix}$$

وبعد ذلك يتم إحلال المتجه مكان كل عمود من أعمدة المصفوفة A ليكون لدينا المصفوفات الجديدة التالية:

$$C = \begin{bmatrix} 6 \\ 30 \\ 8 \end{bmatrix}$$

فعند إحلال C مكان العمود الأول أو الثاني أو الثالث، تكون لدينا المصفوفات A_1 و A_2 و A_3 على التوالي:

$$A_1 = \begin{bmatrix} 6 & 0 & 2 \\ 30 & 4 & 6 \\ 8 & -2 & 3 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 1 & 6 & 2 \\ -3 & 30 & 6 \\ -1 & 8 & 3 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 1 & 0 & 6 \\ -3 & 4 & 30 \\ -1 & -2 & 8 \end{bmatrix}$$

وبالتالي فإن إستخراج قيم المتغيرات X_1 , X_2 , X_3 ستكون بالشكل التالي:

$X_1 = |A_1|/|A| = (-40)/(44) = -10/11$

$X_2 = |A_2|/|A| = (72)/(44) = 18/11$

$X_3 = |A_3|/|A| = (152)/(44) = 38/11$

القيم المميزة والمتجهات المميزة Eigen values & Eigen vectors

في تحليل متعدد المتغيرات، غالباً ما تكون لدينا مصفوفة مربعة $A_{n,n}$ ونحتاج إلى تحديد القيم والمتجهات المميزة لهذه المصفوفة.

وهذا واضح في طرق المكونات الرئيسية Principle components وإستخداماتها في طرق تحليل أخرى وفي مقدمتها التحليل العاملي Factor analysis. ولأننا نتعامل في هذه الطرق مع مصفوفة التباين - التباين المشترك أو مصفوفة معاملات الإرتباط، فإن المصفوفة تكون في العادة مربعة ومتماثلة بنفس الوقت.

ملاحظة:

لاحظ أن التماثل لا يتحقق إلا مع المصفوفة المربعة. ولذلك يمكن الإكتفاء بالقول مصفوفة متماثلة.

ولتوضيح كيفية تحديد القيم والمتجهات المميزة لمصفوفة ما، سنعرض ذلك من خلال المثال التالي:

مثال : لنفترض أنه لدينا مصفوفة التباين - التباين المشترك A التالية:

$$A = \begin{bmatrix} 5 & -2 \\ -2 & 2 \end{bmatrix}$$

وعلينا حل المعادلة:

$$Det(A - \lambda I) = 0.0$$

بالنسبة إلى λ والتي تمثل لنا متجه القيم المميزة للمصفوفة A.

وهذا يعني لنا العمل على:

$$\begin{vmatrix} 5-\lambda & -2 \\ -2 & 2-\lambda \end{vmatrix} = 0.0$$

وبالنتيجة يكون لدينا:

$$(5 - \lambda)(2 - \lambda) - (-2)(-2) = 0.0$$

أو :

$$\lambda^2 - 7\lambda + 6 = 0.0$$

وبحل المعادلة هذه نحصل على:

$$(\lambda - 1)(\lambda - 6) = 0.0$$

أي أنه لدينا: $\lambda = 1$ أو $\lambda = 6$

وهاتان القيمتان هما التعبير عن القيمة المميزة الأولى (الأكبر) $\lambda_1 = 6$ والقيمة المميزة الثانية $\lambda_2 = 1$.

ملاحظة:

يجب أن يبقى في بالنا أن:

$$\lambda_1 \geq \lambda_2$$

و

$$\lambda_1 + \lambda_2 = \text{trace}(A) = 7$$

ولغرض تحديد المتجه المميز مقابل كل قيمة مميزة، فإننا نستخدم الآتي:

$$\begin{pmatrix} 5-\lambda_1 & -2 \\ -2 & 2-\lambda_1 \end{pmatrix}\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 5-6 & -2 \\ -2 & 2-6 \end{pmatrix}\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} -1 & -2 \\ -2 & -4 \end{pmatrix}\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

حيث أن $\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix}$ هو المتجه المميز المقابل للقيمة المميزة λ_1

وبإستخدام أياً من المعادلتين (المتجانستين):

$$-a_{11} - 2a_{12} = 0.0$$

$$-2a_{11} - 4a_{12} = 0.0$$

سنحصل على: $a_{11} = -2a_{12}$

وبالتنسيق مع شرط كون $(a_{11})^2 + (a_{12})^2 = 1$

سنحصل على:

$$\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix}$$

وبتطبيق نفس الإسلوب مع $\lambda_2 = 1$ نحصل على المتجه المميز الثاني المقابل لها وهو:

$$\begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} = \begin{pmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{pmatrix}$$

ملاحظة1:

من المهم هنا ولغرض تطبيق طرق تحليل متعدد المتغيرات والتي تعتمد على هذه القيم، أن تتحقق لدينا حالة التعامدية Orthogonality ما بين هذه الإتجاهات (المتجهات المميزة) وهذا يعني:

1- $\sum_j a_{ij}^2 = 1 \quad , \quad i = 1, 2$

2- $\sum_j a_{1j} a_{2j} = 0.0$

وهذان الشرطان متحققان في مثالنا أعلاه.

ملاحظة2:

في حالة إستخدامنا مصفوفة التباين-التباين المشترك، فإن:

$$\sum_i \lambda_i = \sum_i \mathrm{var}(X_i)$$

أما في حالة إستخدامنا مصفوفة الإرتباط، فإن:

$$\sum_i \lambda_i = p$$

وإن p = عدد المتغيرات ضمن المصفوفة

والمثال أعلاه يمكن أن يتطابق مع الحالة الأولى.

والمثال التالي ينطبق على إستخدام الحالة الثانية:

مثال: لنفترض المصفوفة:

$$A = R = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

أي أنها تمثل مصفوفة الإرتباط بين متغيرين X_1 و X_2 وأن $r_{12} = 0.7$

وهذا يعني لنا العمل على:

$$\begin{vmatrix} 1-\lambda & 0.7 \\ 0.7 & 1-\lambda \end{vmatrix} = 0.0$$

وبالنتيجة يكون لدينا:

$$(1 - \lambda)(1 - \lambda) - 0.49 = 0.0$$

أو:

$$\lambda^2 - 2\lambda + 1 = 0.49$$

وبحل المعادلة هذه نحصل على القيمتان المميزتان:

$\lambda_2 = 0.3$ و $\lambda_1 = 1.7$

ولغرض تحديد المتجه المميز مقابل كل قيمة مميزة، فإننا نستخدم الآتي:

$$\begin{pmatrix} 1-\lambda_1 & 0.7 \\ 0.7 & 1-\lambda_1 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 1-1.7 & 0.7 \\ 0.7 & 1-0.7 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} -0.7 & 0.7 \\ 0.7 & -0.7 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

حيث أن $\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix}$ هو المتجه المميز المقابل للقيمة المميزة λ_1

وهنا نحصل على المتجه المميز الأول : $\begin{pmatrix} a_{11} \\ a_{12} \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$

وفي حالة إستخدامنا القيمة المميزة $\lambda_2 = 0.3$ نحصل على:

$$\begin{pmatrix} 0.7 & 0.7 \\ 0.7 & 0.7 \end{pmatrix} \begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

ومنها سنحصل على المتجه المميز الثاني:

$$\begin{pmatrix} a_{21} \\ a_{22} \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

وهما متعامدان.

مستويات قياس المتغيرات
Variables Levels of Measurment

بطبيعة الحال، عندما نتعامل مع أي نوع من الطرق الإحصائية، فإننا نتعامل مع متغيرات تختلف في تصنيفها من حيث طبيعة القيم التي تشير إليها. ونقصد بذلك، طبيعة مستوى القياس الذي تم بموجبه تحديد قيمة معينة لأي وحدة إحصائية ضمن المتغير الواحد. أي أن كل متغير يخضع لمستوى محدد من القياس تتعامل معه جميع وحداته عند تحديد قيمها. وعندما نستعرض العديد من الطرق الإحصائية سنجد طريقة ما تتطلب متغيرات بمستوى قياس محدد واحد لجميع المتغيرات وطرق أخرى تستوعب متغيرات بأكثر من مستوى قياس واحد.

إن ضرورة توضيح هذا الموضوع هنا هو كون طرق التحليل الإحصائي متعدد المتغيرات التي سنتناولها ضمن هذا الكتاب تتباين أحياناً حسب طبيعة مستويات قياس المتغيرات التي تتعامل معها. وبالتالي فإننا سنتناول بشئ من التفصيل طبيعة مستويات القياس المعروفة هذه مع أمثلة مناسبة لغرض التوضيح. وبذلك سوف نشير إلى طبيعة مستوى القياس لكل متغير يتم استخدامه ضمن أي طريقة تحليل في هذا الكتاب.

وبشكلٍ عام، فإنه من المعروف بأن هنالك أربعة أنواع من مستويات القياس والتي سنذكرها تباعاً حسب القوة (درجة التطور) التي تتميز بها وهي:

المقياس الإسمي (Nominal Scale)

وهو أدنى مستويات القياس ويناسب المتغيرات الكيفية أو النوعية التي تتطلب تصنيف الأفراد إلى مجموعات منفصلة للتمييز بينهم في سمة معينة، ويكون الهدف من عملية القياس في هذه الحالة هو التصنيف الذي يراعي الفروق النوعية بين الأفراد. والأعداد المستخدمة في هذا المستوى من القياس تعد بمثابة رموز بسيطة تستخدم كأسماء لفئات أو مجموعات منفصلة ومتمايزة. التصنيفات في هذه الحالة مختلفة وغير متكررة والأرقام (الأعداد) لم توضع إلا لسهولة التعامل مع المجموعات وليس لها أي دلالة رقمية. فالبيانات (الأعداد) في هذه الحالة فقط تصنّف البيانات ولا تعطي لها أي ترتيب.

ولا نستطيع أن نجري أي عمليات حسابية على هذه الأعداد ولكن بالإمكان عد الأفراد في كل فئة. ومن أمثلة متغيرات هذا المستوى: النوع (ذكور أو إناث) وقد نرمز العدد (1) للذكور والعدد (2) للإناث فنعرف مثلاً أنه لدينا 14 ذكور، 16 إناث بالنسبة لمتغير النوع وهكذا لأمثلة أخرى مثل الجنسية، والديانة، والحالة الاجتماعية، أو حسب مناطق السكن (جنوب، شرق، شمال، غرب)، أو ألوان السيارات (أخضر، أسود، أبيض) فيتم تصنيف ألسيارات بحسب ألوانها ولا يمكن ترتيبها.

1) المقياس الرتبوي (الترتيبي) (Ordinal Scale)

هذا المقياس يصنّف البيانات كما هو حال المقياس السابق لكن يضيف إليها خاصية الترتيب، بحيث أنه يمكن وضع التصنيفات في ترتيب واضح متسلسل. وبذلك يعتبر أكثر تطوراً من المقياس الإسمي. ومن الأمثلة الواضحة على هذا النوع من المقاييس هي المقاييس الخاصة بالتقييم (Rating Scale) أو مقاييس لكرت (Likert Scale) المستخدمة في تصنيف أجوبة أسئلة الإستبانة. فهو يفيد الترتيب بين الأفراد أو وحدات المتغير ولكن ليس من الضروري أن تكون الفروق في مقدار أو درجة الخاصية بين كل رتبتين متجاورتين منتظمة. عندما تكون المشاهدات مختلفة فقط من فئة إلى أخرى، بل فيما يمكن ترتيبها بالنسبة إلى معيار معين فقد يقال عنها بأنها مقاسة بمقياس رتبوي. المرضى في دور النقاهة قد يتم وصفهم (غير متحسنين/ متحسنين/ متحسنين جداً). الأشخاص قد يتم تصنيفهم طبقاً لحالتهم الإقتصادية والإجتماعية مثل (واطئ/متوسط/مرتفع)، ودرجة الذكاء عند الأطفال قد تكون (فوق المعدل/ في المعدل/ تحت المعدل). في كل واحد من هذه الأمثلة نرى أن أعضاء الفئة الواحدة يعتبرون متساوين لكن أعضاء إحدى الفئات يعتبرون أوطأ، أسوأ، أو أصغر من أعضاء الفئة الأخرى وهي التي تحمل تباعاً نفس العلاقة مع فئة ثالثة. فمثلاً المريض المتحسن جداً يكون أكثر صحة من مريض آخر تم تصنيفه كمتحسن بينما المريض المتحسن يكون أحسن حالاً من الآخر غير المتحسن. وبالطبع من المستحيل الإستنتاج أن الفرق بين أعضاء إحدى الفئات وأخرى مجاورة في الأسفل يساوي الفرق بين أعضاء تلك الفئة وأعضاء الفئة المجاورة لها في الأعلى. أي أن درجة التحسن بين غير المتحسن والمتحسن قد لا تكون نفسها بين المتحسن والمتحسن جداً. إن عمل الأرقام المعطاة لبيانات رتبوية هو لترتيب (أو تدريج to rank) المشاهدات من الأوطأ إلى الأعلى وهذا بدوره رتبوي. لهذا لا معنى للعمليات الحسابية في هذا المستوى من القياس، على الرغم من القدرة على إجرائها، وذلك لأن نتائج العمليات الحسابية لا تعكس مقدار الكم للصفة المراد قياسها.

2) المقياس الفئوي (الفتري) (Interval Scale)

المقياس الفئوي (الفتري) يعتبر أكثر تطوراً من المقياس الإسمي أو الرتبوي لأنه مع هذا المقياس ليس بالإمكان ترتيب القياسات فقط، لكن المسافة بين أي قياسين تكون معروفة. فنحن نعرف مثلاً، أن المسافة بين القياس 20 والقياس 30 يكون مساوياً للمسافة بين القياس 30 والقياس 40. والمقدرة في عمل هذا، يقود إلى إستخدام وحدة المسافة ونقطة الصفر وكلاهما عفوي. فنقطة الصفر المختارة ليست صفراً حقيقياً لكونها لا تشير إلى الغياب الكلي للمقدار الذي يكون مقاساً. فالفروق أو المسافات المتساوية على هذا المقياس متساوية تدل على مقادير متساوية من الخاصية التي نقيسها، ولذا يمكن جمع هذه المسافات أو طرحها أو ضربها مع مراعاة أنه لا يوجد لمقياس المسافة صفر حقيقي أو مطلق (يدل على إنعدام الشيء أو عدم

وجود الخاصية). وربما أحسن مثال لمقياس الفترة هو الطريقة التي تقاس بها درجة الحرارة. فوحدة القياس هي الدرجة ونقطة المقارنة هي " درجة الصفر" المختارة عفوياً. فدرجة الصفر مئوي لا تعني إنعدام الحرارة من الوجود، كما أن الفرق ما بين درجتي الحرارة 25 و 28 هو نفسه الفرق ما بين درجتي الحرارة 81 و 84. ومن جانب آخر نجد أن درجة الحرارة الصفر بمقياس الفهرنهايت هي 32 ، كما أن الصفر الجامعي في المساق هو 35. إن مقياس الفترة يختلف عن المقياس الإسمي والمقياس الرتبوي بكونه مقياس مقادير حقيقية ويستخدم للبيانات الكمية.

3) المقياس النسبي (Ratio Scale)

يحتفظ هذا النوع من المقاييس بمزايا الثلاثة أنواع السابقة، فهو يصنّف، يرتّب ويوضح المسافات بشكل متساوي وموزون. وبالإضافة لذلك، فإن الخاصية التي ينفرد بها مقياس النسب هي نقطة الصفر الحقيقية أي صفرمطلق يناظر بالفعل إنعدام الخاصية والسمة المقاسة. ويمكن إجراء جميع العمليات الحسابية الأساسية على هذه المقاييس ومن أمثلتها مقاييس الوزن، والحجم، والطول والمسافة وغيرها. وسمي بمستوى القياس النسبي، لأن نسبة الأرقام إلى بعضها ذات معنى ودلالة. العمليات الرياضية (الحسابية) والمقارنات عند هذا المقياس لها معنى ويمكن للباحثين إجراء عمليات القسمة والضرب دون تغيير في الخصائص. وتمثل المسطرة مثالاً بسيطاً للمقياس النسبي، فالفرق فيها بين نقاط القياس متساو في العرض، وهناك نقطة صفر حقيقية على المسطرة التي تجعل أي قياس تحت الصفر ليس بذي معنى. ولذلك يمكن تصنيفه بأنه أعلى مستوى للقياس. وهذا المقياس ينطبق دائماً على قياس المتغير الكمي والذي سيتم تناوله تالياً. وبذلك قد نطلق عليه "المقياس الكمي".

وإنطلاقاً من هذا الإستعراض لمستويات القياس للمتغيرات، نلاحظ أن المتغيرات يمكن أن تكون، بحسب مستوى قياسها، مترية (كمية)، مثل (درجات الامتحان، السرعة، الفئات العمرية، مستويات الدخل،...) وهذه تشمل المقياسين (النسبي Ratio) و (الفئوي Interval)، أو لامترية (نوعية) مثل (اللون، الجنس، مستوى التعليم، الرتبة العسكرية....) وهذه تشمل المقياسين (الإسمي Nominal) و (الرتبوي Ordinal).

كما ينقسم المتغير الكمي إلى متغير متصل (مستمر Continuous) والذي يمكن أن تأخذ مقاديره أي قيمة، ومتغير منفصل (متقطع Discrete) تنحصر مقاديره في قيم العد (الأعداد الصحيحة). وتصنيف المتغير الكمي إلى متصل أو منفصل يعتمد على أداة القياس المستخدمة في قياسه وليس على القيم التي يظهر بها ونتعامل بموجبها. ومثالاً على ذلك قيم درجات الحرارة حيث أننا نتعامل في العادة مع أعداد صحيحة منفصلة (1 ، 2 ، 3 ،.......) ويظهر ذلك بوضوح عند مراقبة نشرات الأنواء الجوية. ولكن مقياس درجة الحرارة لا ينتقل

من الدرجة (1) إلى الدرجة (2) مباشرةً وإنما يقرأ أي جزء ما بين الدرجتين، وبالتالي فإن درجة الحرارة هي متغير متصل ولو تعاملنا مع قيمه المنفصلة.

ملاحظة:

من خلال إستعراضنا لمستويات القياس هذه، والعلم بأن غالبية الطرق الإحصائية في التحليل المتعدد التي سيتم تناولها في هذا الكتاب تعتمد مصفوفة التباين أو مصفوفة معامل الإرتباط في العمليات الحسابية الخاصة بها والتي تتعامل حصراً مع المتغيرات الكمية، يصبح من الواضح جداً بأن المقياس النسبي (Ratio scale) هو المقياس المناسب لجميع المتغيرات التي تدخل في أي مصفوفة وقد يكون من الممكن التعامل مع المقياس الفئوي (Interval scale) وسوف نراعي ذكر ذلك عند تناول هذه الطرق.

تحليل الإنحدار والإرتباط المتعدد
Multiple Regression & Correlation Analysis

أولا: تحليل الإنحدار الخطي البسيط

عند دراسة العلاقة بين المتغيرات فإن الخطوة الأولى تبدأ في تحديد المتغيرات التي تدخل في هذه العلاقة. فإذا كانت هذه العلاقة بسيطة أي بين إثنين من المتغيرات، ففي هذه الحالة عادة ما نفكر بأحد المتغيرين بأنه المتغير السببي ويوصف بأنه المتغير المستقل (x) والمتغير الآخر بأنه المتغير التابع أو متغير الاستجابة (y) أي أن y هو دالة x لكن القيمة المشاهدة إلى y لا يمكن أن ترتبط بعلاقة خطية مضبوطة مع القيمة المشاهدة إلى x في كل محاولة من المحاولات. لذلك نلجأ إلى إضافة حد جديد يسمى حد الخطا أو الخطأ العشوائي الذي يمثل الفشل بالنسبة للقيمة المشاهدة إلى y في أن تكون مساوية للعلاقة الخطية ($\beta_0 + \beta_1 x$) فتصبح الصيغة الملائمة بالشكل التالي:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1 , 2 , ..., n$$

حيث أن:

y_i : تمثل متغير الاستجابة (المعتمد) في المشاهدة (i) وهو متغير كمي (نسبي أو فئوي).

x_i : تمثل المتغير التوضيحي (المستقل) في المشاهدة (i) وهو متغير كمي (نسبي أو فئوي).

e_i : يمثل حد الخطأ في المشاهدة (i).

الفرضيات الخاصة بنموذج الإنحدار البسيط

وأهمها الفرضيات الخاصة بحد الخطأ e_i والتي تشمل الآتي:

- e_i متغير عشوائي
- يتوزع توزيع طبيعي
- وسطه الصفر $E(e_i)=0$
- تباينه ثابت σ^2 لكل قيم x_i
  أي أن $e_i \sim N(0 , \sigma^2)$
- المتغير العشوائي e_i مستقل عن قيم x_i أي أن
  $$cov(x_i , e_i) = 0.0$$
- الخطأ العشوائي في أي محاولة يكون مستقلاً عن الأخطاء العشوائية في محاولة أخرى أي أن:
  $$cov(e_i , e_j) = 0.0 \qquad i \neq j$$

أما الفرضيات الخاصة بالمتغير المستقل x_i فتنحصر عموماً بكون قيمه ثابتة.

<u>إستنتاجات خاصة بالمتغير المعتمد</u> (y)

- بما أن $E(e_i) = 0$ فإن $E(y_i) = \beta_0 + \beta_1 x_i$
- بما أن $v(e_i) = \sigma^2$ لكل قيم i فإن $v(y) = \sigma^2$ لكل قيم i
- y_i متغير عشوائي يتبع التوزيع الطبيعي بمتوسط مقداره $\beta_0 + \beta_1 x_i$ وتباين ثابت مقداره σ^2. أي أن: $y_i \sim N(\beta_0 + \beta_1 x_i , \sigma^2)$
- بما أن قيم الاخطاء العشوائية غير مرتبطة ببعضها البعض، لذلك فإن y_i و y_j غير مرتبطة أيضاً، أي أن: $Cov(y_i,y_j) = 0$

تقدير دالة الإنحدار

لإيجاد معادلة الإنحدار التقديرية لابد من إيجاد تقديرات معاملات الإنحدار (β_0 ، β_1) وذلك بإستخدام إحدى طرق التقدير المعروفة والتي تعطي نتائج تقديرية للمعالم تحمل الكثير من الصفات المرغوبة في التقديرات الاحصائية، ومن هذه الطرق:

- طريقة المربعات الصغرى Least Squares Method
- طريقة الإمكان الأعظم Maximum Liklihood

وسوف نقتصر هنا على إستخدام الطريقة الأولى وهي طريقة المربعات الصغرى العادية ordinary least squares.

إن أساس طريقة المربعات الصغرى يعتمد على تقدير قيم المعالم المجهولة لنموذج الإنحدار β_0 و β_1 والتي تجعل مجموع مربعات الأخطاء العشوائية في نهايتها الصغرى.

وبإتباع إجراءات النهايات الصغرى يمكن حساب التقديرات b_0 و b_1 وذلك باستخدام أسلوب التفاضل الجزئي بالنسبة إلى كل من B_0 و B_1 ثم نجعل هذه المشتقات الجزئية مساوية للصفر. وبحل المعادلات الآنية بالنسبة إلى β_0 و β_1 نحصل على ما يلي:

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

ولغرض تحليل التباين ومعرفة معنوية نموذج الإنحدار الخطي البسيط يمكننا إستخدام العلاقات الآتية:

مجموع المربعات الكلي: مجموع مربعات الإنحدار + مجموع مربعات الخطأ البسيط.

أي أن:

$$SST = SSR + SSE$$

حيث أن:

$$SST = \sum (y_i - \bar{y})^2 = \sum y^2{}_i - n\bar{y}^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = \sum \hat{y}^2{}_i - n\bar{y}^2 = b_1(\sum x_i y_i - n\bar{x}\bar{y})$$

$$SSE = \sum (y_i - \hat{y}_i)^2 = SST - SSR$$

وبذلك فإن جدول تحليل التباين سوف يكون كالآتي:

S.V.	df	SS	MS	F
Regression	1	SSR = $b_1(\sum x_i y_i - n\bar{x}\bar{y})$	MSR = SSR/1	MSR/MSE
Error	n-2	SSE = by Sub.	MSE = SSE/(n-2)	
Total	n-1	SST = $\sum y^2{}_i - n\bar{y}^2$		

ومن جدول تحليل التباين أعلاه يمكن أن يتضح لنا مدى معنوية نموذج الإنحدار أو هل أن المتغير المستقل له تأثير معنوي (جوهري) على المتغير المعتمد وذلك بإختبار الفرضية التالية:

$H_0 : \beta_1 = 0.0$

$H_1 : \beta_1 \neq 0.0$ مقابل الفرضية البديلة:

ويتم ذلك بإستخدام اختبار F والذي تتضح قيمته في الجدول أعلاه.

أو إستخدام اختبار t حسب الصيغة التالية:

$$T = b_1/\sqrt{\text{Var(b1)}}$$

Coefficient of Determination **معامل التحديد**

يمكننا الحصول على معامل التحديد الذي يمثل نسبة الإنحرافات الكلية الموضحة أو المشروحة بواسطة معادلة الإنحدار التقديرية أو نسبة مساهمة معادلة الإنحدار التقديرية في تفسير أو شرح الإنحرافات الكلية عن قيم y حول الوسط الحسابي $\bar{y}$ وهو يأخذ الصيغة التالية:

$$R^2 = SSR/SST = \sum (\hat{y}_i - \bar{y})^2 / \sum (y_i - \bar{y})^2 \quad , \quad 0 \leq R^2 \leq 1$$

ثانياً: نموذج الإنحدار الخطي المتعدد

إن نموذج الإنحدار الخطي البسيط يقتصر فقط على تحليل العلاقة بين متغيرين أحدهما مستقل والآخر معتمد ولكن في الواقع أن أغلب الظواهر الإقتصادية والإجتماعية وغيرها بحاجة إلى وضع المتغير المعتمد كدالة لأكثر من متغير مستقل. ففي مثل هذه العلاقات يجب تمثيلها بنماذج خطية متعددة وأن المعادلة التي تمثل هذه العلاقة تنتمي لنموذج الإنحدار الخطي المتعدد وفيه نفترض وجود علاقة خطية بين أحد المتغيرات y (المتغير المعتمد) وعدد k من المتغيرات المستقلة ($X_1, X_2 ....X_k$) والتي تدعى بالمتغيرات التوضيحية بسبب إستخدامها في توضيح التباعد في المتغير y. وكل هذه المتغيرات تظهر بمقياس كمي نسبي في الغالب ولكن يمكن التعامل أيضاً مع المقياس الفئوي. بالإضافة إلى ذلك، فإن بعض المتغيرات المستقلة يمكن أن تظهر بمقياس رتبوي أو حتى إسمي ويطلق عليها عندئذٍ بالمتغيرات الوهمية Dummy Variables وهذا سيضيف بعض الصعوبات في تفسير النتائج. وسنتناول مثالاً بهذه الحالة نهاية هذا الفصل لتوضيح طريقة تحليل النتائج المرتبطة بها. ويمكن صياغة نموذج الإنحدار الخطي المتعدد بالشكل التالي:

$$y = f[(x_1 , x_2 , ..., x_k) , e]$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + e_i \quad , \; i = 1 , 2 , ..., n$$

حيث نلاحظ وجود عدد k من المتغيرات المستقلة و (k+1) من المعالم و n من المشاهدات.

ولغرض تبسيط عرض النموذج الخطي المتعدد سوف نستخدم اسلوب المصفوفات في عرض مشاهدات المتغيرات المستقلة والمتغير المعتمد إضافة إلى إستخدام هذا الأسلوب عند تقدير معالم النموذج الخطي المتعدد. لذلك يمكن التعبير عن النموذج أعلاه وكما يلي:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

حيث أن:

Y قيمة مشاهدات المتغير المعتمد وهو متجه ذو رتبة (n,1).

X مصفوفة مشاهدات المتغيرات المستقلة ذو رتبة (n,k+1).

β قيمة لمعالم النموذج الخطي المتعدد المطلوب تقديرها وهو متجه ذو رتبة (k+1,1).

e قيمة للأخطاء وهومتجه ذو رتبة (n,1).

ولغرض التفكير بأسلوب ملائم لتقدير معالم النموذج يجب علينا النظر في الفروض الأساسية الخاصة بمتجه الأخطاء العشوائية والتي يفترض أن تنطبق على جميع المشاهدات أي أن:

$$e_i \sim N(0, \sigma^2)$$

وهي أن قيم الأخطاء تتوزع توزيعاً طبيعياً بتوقع (متوسط) قدره:

$$E(e) = E\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} E(e_1) \\ E(e_2) \\ \vdots \\ E(e_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

وتباين ثابت $\sigma^2 I_n$ ويمكن وضعه بالصيغة التالية:

$$Var(e) = E(e.e') = E(\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix})$$

$$= E\begin{bmatrix} e_1^2 & e_1e_2 & \cdots & e_1e_n \\ e_2e_1 & e_2^2 & \cdots & e_2e_n \\ \vdots & \vdots & \vdots & \vdots \\ e_ne_1 & e_ne_2 & \cdots & e_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(e_1^2) & E(e_1e_2) & \cdots & E(e_1e_n) \\ E(e_2e_1) & E(e_2^2) & \cdots & E(e_2e_n) \\ \vdots & \vdots & \vdots & \vdots \\ E(e_ne_1) & E(e_ne_2) & \cdots & E(e_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} = \sigma^2 I_n$$

وذلك لأنه، وحسب الإفتراض بأن التباين ثابت، أي:

$$\sigma_1^2 = \sigma_2^2 = \dots\dots\dots = \sigma_n^2 = \sigma^2$$

بالإضافة إلى ذلك هناك فروض أخرى يجب توفرها في نموذج الإنحدار الخطي المقصود هي:

1) عدم وجود علاقة خطية محددة أو تامة بين المتغيرات المستقلة.
2) يجب أن يكون عدد المشاهدات أكبر من عدد المتغيرات المستقلة.

تقدير معالم نموذج الإنحدار الخطي المتعدد

سنقوم بإستخدام طريقة المربعات الصغرى لتقدير معالم النموذج لأنها تهدف إلى جعل مجموع مربعات الأخطاء أقل ما يمكن. (الرموز في أدناه هي مصفوفات مثل X والبقية متجهات):

$$y = X\beta + e$$

$$e = y - \hat{y} = y - Xb$$

$$e'e = (y - Xb)'(y - Xb)$$

$$= y'y - y'Xb - b'X'y + b'X'Xb$$

$$= y'y - 2\,b'X'y + b'X'Xb$$

$$\frac{\partial e'e}{\partial b} = -2\,X'y + 2\,X'Xb = 0.0$$

وبذلك نحصل على:

$$X'y = X'Xb$$

$$b = (X'X)^{-1}X'y$$

ويمكننا إيجاد مكونات العلاقة السابقة عن طريق إجراء العمليات الإعتيادية على المصفوفات وكما يلي:

$$X'X = \begin{bmatrix} n & \sum x_1 & \sum x_2 & \cdots & \sum x_k \\ \sum x_1 & \sum x_1^{2} & \sum x_1x_2 & \cdots & \sum x_1x_k \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_k & \sum x_kx_1 & \sum x_kx_2 & \cdots & \sum x_k^2 \end{bmatrix}$$

$$X'y = \begin{bmatrix} \sum y \\ \sum x_1 y \\ \sum x_2 y \\ \vdots \\ \sum x_k y \end{bmatrix}$$

سمات مقدرات طريقة المربعات الصغرى

وتمتاز مقدرات المربعات الصغرى بصفات كثيرة إلا أن أهمها:

1- عدم التحيز unbiasedness أي أن:

$$E(\hat{\beta}) = E(b) = \beta$$

وهذا واضح من خلال حقيقة كون:

$$\hat{\beta} = b = (X'X)^{-1} X'y$$
$$E(y) = X\beta$$

وبالتالي يكون لدينا:

$$\begin{aligned} E(\hat{\beta}) &= E((X'X)^{-1} X'y) \\ &= ((X'X)^{-1} X'E(y)) \\ &= (X'X)^{-1} X'X\beta \\ &= \beta \end{aligned}$$

2- أقل تباين minimum variance

لكي نثبت أن التقديرات بطريقة المربعات الصغرى لها أقل تباين، لابد أن نقوم أولاً بتقدير التباين للمتجه $b = \hat{\beta}$ وكما يلي:

$$\begin{aligned} var(\hat{\beta}) = var(b) &= E(\hat{\beta} - \beta)^2 \\ &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\ &= E[(X'X)^{-1} X'e][e'X((X'X)^{-1}] \\ &= (X'X)^{-1} X'E(e\,e')X((X'X)^{-1} \\ &= \sigma^2 I_n (X'X)^{-1} X\,X\,(X'X)^{-1} \\ &= \sigma^2 I_n (X'X)^{-1} \end{aligned}$$

وبذلك يكون توزيع المتجه $\hat{\beta} = b$ هو طبيعي وكما يلي:

$$\hat{\beta} \sim N(\beta\,,\ \sigma^2 I_n (X'X)^{-1})$$

ولكي نتأكد من أن هذا التقدير لتباين $\hat{\beta}$ هو الأقل، سنلاحظ ذلك من خلال الآتي:

لنفترض أن b^* تمثل أى تقدير أخر غير متحيز للمتجه β حصلنا عليه بطريقة أخرى غير طريقة المربعات الصغرى:

$$b^* = [(X'X)^{-1} X' + D]y$$

حيث أن D مصفوفة ثوابت وبذلك من الواضح أن يكون:

$$Var(b^*) = [(X'X)^{-1} X' + D]var(y) [(X'X)^{-1} X' + D]'$$

$$= \sigma^2 I_n \; [(X'X)^{-1} + DX\,(X'X)^{-1} + (X'X)^{-1} X' D' + DD']$$

وحتى لو إفترضنا أن DX = 0.0 فإن ذلك يبقى مساوياً إلى:

$$= \sigma^2 I_n \; [(X'X)^{-1} + DD']$$

وهذا يثبت أن تقديرات المربعات الصغرى لها أقل تباين.

تحليل التباين لنموذج الإنحدار الخطي المتعدد

سبق وأن أوضحنا في الإنحدار الخطي البسيط أن مجموع مربعات الإنحرافات الكلية يمكن أن يجزأ إلى قسمين هما مجموع مربعات الإنحدار ومجموع مربعات الأخطاء حيث:

SST = SSR + SSE

SST = $y'y - n\bar{y}^2$

SSR = $\hat{y}'\hat{y} - n\bar{y}^2$ = (xb)'xb - $n\bar{y}^2$

= b'x'x(x'x)$^{-1}$x'y - $n\bar{y}^2$

= b'x'y - $n\bar{y}^2$

SSE = SST – SSR

= y'y - b'x'y

وبذلك يكون جدول تحليل التباين كما يلي:

S.V.	df	SS	MS	F
Regression Error	k-1 n-k	SSR = b'x'y - $n\bar{y}^2$ SSE = y'y - b'x'y	MSR = SSR/(k-1) MSE = SSE/(n-k)	MSR/MSE
Total	n-1	SST = $y'y - n\bar{y}^2$		

إختبار معنوية الإنحدار

يمكننا أن نميز نوعين من الإختبارات:

1. إختبار معنوية الإنحدار في حالة النموذج الكامل (جميع المعلمات بضمنها β_0):

H$_0$: $\beta = 0$

H$_1$: $\beta \neq 0$

وتقارن قيمة F = $\dfrac{\hat{\beta}'x'x\hat{\beta}/k}{SSE/(n-k)}$ مع $f(k, n-k)$ الجدولية.

2. إختبار معنوية الإنحدار في حالة النموذج المختزل (بدون β_0):

H0: $\beta_1 = \beta_2 = ... = \beta_k = 0$

H1: $\beta_i \neq 0 \quad , \; i = 1,2,......,k$

وتُقارن قيمة $F = \frac{\hat{\beta}' x' x \hat{\beta} / (k-1)}{SSE / (n-k)}$ مع $f(k-1, n-k)$ الجدولية وهو في العادة ما يعتمد لأن β_0 لا تشكل شيئاً مهماً في مساهمة المتغيرات المستقلة في النموذج.

معامل التحديد R^2

لاحظنا سابقا أن نموذج الإنحدار الخطي المتعدد يمثل العلاقة بين المتغير المعتمد وعدد من المتغيرات المستقلة. كما أن مجموع المربعات الكلي لهذا النموذج يتكون من مجموع مربعات الإنحدار للمتغيرات المستقلة مضافاً إليها مجموع مربعات الأخطاء (غير المُفسَرة). وإذا إفترضنا أن معادلة الإنحدار تمثل هذه العلاقة تمثيلاً جيداً فإنه من الضروري أن تكون نسبة مجموع مربعات الإنحدار إلى مجموع المربعات الكلي كبيرة وهذا ما يسمى بمعامل التحديد R^2 والذي يمثل نسبة مساهمة المتغيرات المستقلة في تفسير التباين في المتغير المعتمد Y وحسب الصيغة التالية:

$$SST = SSR + SSE$$

$$1 = SSR/SST + SSE/SST$$

$$R^2 = 1 - SSE/SST$$

وبالتالي فإن:

$$R^2 = SSR/SST = \frac{b'x'y - n\bar{y}^2}{y'y - n\bar{y}^2}$$

معامل التحديد المصحح (adjusted R^2)

يمتاز معامل التحديد R^2 بأنه لو أضيف متغير مستقل إلى النموذج فان قيمته سترتفع حتى وإن لم يكن للمتغير المضاف من الأهمية ما يستحق معها إدخاله في النموذج. ولذا ولغرض الحصول على معيار أفضل لقياس مدى قابلية مجاميع مختلفة من المتغيرات لتحليل العلاقة قيد الدراسة وبنفس الوقت الأخذ بنظر الإعتبار عدد المتغيرات المشمولة، فإنه يتم حساب ما يسمى بمعامل التحديد المصحح بموجب الصيغة التالية:

$$\bar{R}^2 = 1 - \frac{\sum e_i^2 / (n-k)}{\sum (y_i - \bar{y})^2 / (n-1)}$$

وبذلك فإن العلاقة بين معامل التحديد المصحح ومعامل التحديد غير المصحح هي:

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}\left(1 - R^2\right)$$

ويلاحظ أن قيمة $\overline{R}^2$ سوف تنخفض عند إضافة متغير مستقل إذا لم تؤدي هذه الإضافة إلى تقليص قيمة ($1-R^2$) بما يعوض عن الزيادة التي تحصل في $\frac{n-1}{n-k}$ نتيجة لإرتفاع قيمة K.

وبعبارة أخرى، من الأفضل عدم إضافة متغير إلى النموذج إذا تسببت إضافته إلى تخفيض قيمة $\overline{R}^2$.

الإرتباط The Correlation

كما سبق وذكرنا في فرضيات نموذج الإنحدارالخطي المتعدد أنه يجب أن تكون العلاقة بين المتغيرات التوضيحية معدومة حتى لا تنتج لدينا مشكلة تعدد الإرتباط الخطي Multicolinearityفي الإنحدار. وبالمقابل، يجب أن تكون هناك علاقة قوية بين المتغيرات التوضيحية من جهة والمتغير المعتمد من جهة أخرى لكي تكون معادلة الإنحدار الخطي المتعدد كفوءة في تفسير الظاهرة المدروسة.

ولدراسة هذه الإرتباطات يمكن تحديد نوعين أساسيين هما:

1. إرتباط المتغير المستقل X مع المتغير المعتمد y وهو ما يسمى بالإرتباط البسيط.
2. إرتباط المتغيرات المستقلة بمجموعها مع المتغير المعتمد وهو ما يسمى بالإرتباط الخطي المتعدد.

ويمكن التعبير عن الإرتباط الخطي البسيط بالصيغ التالية:

$$\begin{aligned} r_{xy} &= \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} \\ &= \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} \\ &= \frac{Cov(x, y)}{\sqrt{V(x)V(y)}} \end{aligned}$$

معامل الإرتباط الجزئي

يمثل معامل الإرتباط الجزئي صافي الإرتباط بين المتغير المعتمد والمتغير المستقل بعد حذف التأثير المشترك لباقي المتغيرات المستقلة على كل من المتغير المعتمد والمتغير المستقل (أي بعد تثبيت المتغيرات الأخرى). مثلا :

$r_{yx_1.x_2}$ أو $r_{y1.2}$ يعني الإرتباط الجزئي بين Y و X_1 بعد حذف تأثير X_2 على كل من X_1 و Y , ويستخدم لتحديد الأهمية النسبية للمتغيرات المستقلة المختلفة في الإنحدار المتعدد علماً أن قيمته، مثلما هي لأي معامل إرتباط، تنحصر ما بين (1+ و 1-) ويأخذ إشارة المعلمة المناظرة ويتم إحتسابه بموجب الصيغة التالية:

$$r_{y1.2} = \frac{r_{y1} - r_{12} r_{y2}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{y2}^2\right)}}$$

أو يمكن حسابه من جدول تحليل التباين بالصيغة:

$$\left| r_{y1.2} \right| = \left| \sqrt{\frac{SSR(X_1 \mid X_2)}{SST - SSR(X_2)}} \right|$$

أما الإرتباطات الجزئية مع إضافة متغيرات مستقلة أخرى فتحتسب معاملاتها على نفس النمط وكما يلي:

في حالة ثلاثة متغيرات مستقلة تكون:

$$r_{y1.23} = \frac{r_{y1.3} - r_{12.3} r_{y2.3}}{\sqrt{\left(1 - r_{12.3}^2\right)\left(1 - r_{y2.3}^2\right)}} = \frac{r_{y1.2} - r_{13.2} r_{y3.2}}{\sqrt{\left(1 - r_{13.2}^2\right)\left(1 - r_{y3.2}^2\right)}}$$

أو

$$\left| r_{y1.23} \right| = \left| \sqrt{\frac{SSR(X_1 \mid X_2, X_3)}{SST - SSR(X_2, X_3)}} \right|$$

وفي حالة أربعة متغيرات مستقلة تكون:

$$r_{y1.234} = \frac{r_{y1.34} - r_{12.34} r_{y2.34}}{\sqrt{\left(1 - r_{12.34}^2\right)\left(1 - r_{y2.34}^2\right)}}$$

أو

$$\left| r_{y1.234} \right| = \left| \sqrt{\frac{SSR(X_1 \mid X_2, X_3, X_4)}{SST - SSR(X_2, X_3, X_4)}} \right|$$

وبصورة عامة فأن معامل الإرتباط الجزئي بين المتغيرين (i و j) بعد جعل جميع تأثيرات المتغيرات الأخرى ثابتةً هو:

$$r_{ij.(all\,other\,variables)} = \frac{-C_{ij}}{\sqrt{C_{ii} C_{jj}}}$$

حيث أن C_{ii} و C_{jj} و C_{ij} هي عناصر معكوس مصفوفة معامل الإرتباط.

ولأجل توضيح ذلك، لنفترض ثلاثة متغيرات مستقلة X_1 , X_2 , X_3 إلى جانب المتغير المعتمد y. في هذه الحالة، تكون لدينا مصفوفة الإرتباط:

$$R = \begin{vmatrix} 1 & r_{12} & r_{13} & r_{1y} \\ r_{21} & 1 & r_{23} & r_{2y} \\ r_{31} & r_{32} & 1 & r_{3y} \\ r_{y1} & r_{y2} & r_{y3} & 1 \end{vmatrix}$$

ولذلك فإن معكوس هذه المصفوفة:

$$R^{-1} = \begin{vmatrix} C_{11} & C_{12} & C_{13} & C_{1y} \\ C_{21} & C_{22} & C_{23} & C_{2y} \\ C_{31} & C_{32} & C_{33} & C_{3y} \\ C_{y1} & C_{y2} & C_{y3} & C_{yy} \end{vmatrix}$$

وبذلك، وعلى سبيل المثال، فإن:

$$r_{23.1y} = \frac{-C_{23}}{\sqrt{C_{22}C_{33}}}$$

أما بالنسبة لمعامل الإرتباط المتعدد والذي يرمز له برمز $R_{y.12......k}$ فإنه يقيس مدى قرب سطح الإنحدار من النقاط المشاهدة، وهذا يعني أن معامل الإرتباط المتعدد يقيس التأثير المشترك لجميع المتغيرات المستقلة على المتغير التابع. ويُعرًف معامل الإرتباط المتعدد على النحو التالي:

$$R_{y.12......k} = \sqrt{\frac{SSR}{SST}} = \sqrt{\frac{b'x'y - n\bar{y}^2}{y'y - n\bar{y}^2}}$$

أيضاً يمكن كتابته (في حالة متغيرين مستقلين وثلاثة متغيرات مستقلة على سبيل المثال) كما يلي:

$$R_{y.12} = \left[1 - \left(1 - r_{y1}^2\right)\left(1 - r_{y2.1}^2\right)\right]^{\frac{1}{2}}$$

ويمكن عمل الاختبارات التالية:

1. بالنسبة للإرتباط المتعدد:

$$H_0:\ r_{y.123\,......\,p} = 0$$

$$H_1:\ r_{y.123\,......\,p} > 0$$

ونرفض H_0 إذا كان:

$$\left(\frac{r^2_{y.123\,....\,p}}{1-r^2_{y.123\,....\,p}}\right)\left(\frac{n-p}{p-1}\right) > f_{\alpha,\,p-1,\,n-p}$$

2. بالنسبة للإرتباط الجزئي (حالة ثلاثة متغيرات مستقلة):

$$H_0:\ r_{y2.13} = 0$$

$$H_1:\ r_{y2.13} \neq 0$$

بإستخدام إحصاءة الإختبار:

$$T = \frac{r_{y2.13}}{\sqrt{\dfrac{1-r^2_{y2.13}}{n-3}}} \sim t_{n-3}$$

إختيار أفضل معادلة إنحدار

إن أحد أصعب مسائل تحليل الإنحدار هي إختيار مجموعة المتغيرات المستقلة المتضمنة في النموذج. وهناك عدد من الطرق التي تساعد في إيجاد أفضل مجموعة من المتغيرات المستقلة. وهذه الطرق بصورة عامة يفضل استخدامها مع الحاسب الآلي لأنها تحتاج إلى عمليات حسابية مطولة جداً وخاصة في حالة وجود عدد كبير من المتغيرات المستقلة. ومن هذه الطرق ما يلي:

1. طريقة الخطوات المتسلسلة Stepwise
2. طريقة كل الإنحدارات الممكنة All possible regression.
3. طريقة الإختيار الأمامي أو المباشر Forward method.
4. طريقة الحذف المعاكس Backward method.

وسنشرح هنا طريقة الإختيار الأمامي أو المباشر Forward method وتتلخص بما يلي:

نبدأ المعادلة بدون أي متغير مستقل ثم نختار المتغيرات المستقلة التي تدخل للمعادلة واحداً بعد الآخر ونتوقف عن الإختيار عندما تقل قيمة F الجزئية عن قيمة f الجدولية المقابلة.

وأول المتغيرات الذي يدخل المعادلة هو المتغير الذي له أعلى قيمة F محسوبة وتزيد عن قيمة f الجدولية. المتغير الثاني الذي يدخل المعادلة أعلاه هو المتغير الذي له أعلى قيمة F جزئية بوجود المتغير الأول المنتخب بالخطوة الأولى وتزيد عن قيمة f الجدولية المعينة لتلك الخطوة. وهكذا نستمر بإضافة المتغير الذي له أعلى قيمة F جزئية وتزيد عن f الجدولية إلى أن نصل إلى أعلى قيمة F جزئية تقل عن f الجدولية فعند ذلك نتوقف عن الإضافة. وستطبق هذه الطريقة على المثال التالي ببيانات حقيقية لتجربة مختبرية: (1)

المشاهدات	Y	X_1	X_2	X_3	X_2
1	78.5	7	26	6	60
2	74.5	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	59	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.5	11	66	9	12
13	109.4	10	68	8	12

والذي يمثل تجربة مختبرية لدراسة تأثير بعض العناصر الفلزية على درجة التوصيل الحراري لقضيب معدني.

حيث أن:

Y تمثل الحرارة المنبعثة

X_1 كميات الومينات الكالسيوم الثلاثي

X_2 كميات سليكات الكالسيوم الثلاثي

X_3 الومينات الكالسيوم الرباعي الحديدي

X_4 كمية سليكات الكالسيوم الثنائي

وجميع هذه المتغيرات بقياس كمي نسبي.

الخطوة الاولى:

أ- نحسب قيمة F لانحدار Y على كل من X_1 و X_2 و X_3 و X_4 وبشكلٍ منفرد كما في الجداول التالية:

تحليل التباين لإنحدار Y على X_1

S.V	DF	SS	MS	F
$R(X_1)$	1	1950.0769	1950.0769	
$ERROR(X_1)$	11	1265.6667	115.0629	**12.60**
TOTAL	12	2715.7631		

تحليل التباين لإنحدار Y على X_2

S.V	DF	SS	MS	F
$R(X_2)$	1	1809.9268	1809.9268	
$ERROR(X_2)$	11	906.3363	82.3942	**21.96**
TOTAL	12	2715.7631		

تحليل التباين لإنحدار Y على X_3

S.V	DF	SS	MS	F
$R(X_3)$	1	776.3626	776.3626	
$ERROR(X_3)$	11	1939.9005	176.3091	**4.40**
TOTAL	12	2715.7631		

تحليل التباين لإنحدار Y على X_4

S.V	DF	SS	MS	F
$R(X_4)$	1	1831.8962	1831.8962	
$ERROR(X_4)$	11	883.8669	80.3515	**22.80**
TOTAL	12	2715.7631		

ب- إن أول متغير يُنتخب ليدخل المعادلة هو X_4 (كمية سليكات الكالسيوم الثنائي) لأنه أعلى قيمة F وهي تزيد عن قيمة:

$$F=22.80 > \quad f_{0.01\,,\,(1,11)} = 3.23$$

أي أن كمية سليكات الكالسيوم الثنائي له التأثير الأوضح على درجة التوصيل الحراري.

ملاحظة:

لو كانت قيمة أعلى F أقل من قيمة f الجدولية هذه، فعندئذ نتوقف ونقول بأنه لا يوجد أي متغير مستقل له أي تأثير معنوي على معدل Y.

الخطوة الثانية:

أ- ولإنتخاب المتغير الثاني لإدخاله في المعادلة التي تضم المتغيرالمنتخب الأول X_4، نحسب قيمة F الجزئية لكل متغير آخر بوجود المتغير X_4. أي نحسب قيم F الجزئية التالية:

$$F(X_1 \mid X_4) = \frac{MSR(X_1 \mid X_4)}{MSE(X_1, X_4)}$$

$$F(X_2 \mid X_4) = \frac{MSR(X_2 \mid X_4)}{MSE(X_2, X_4)}$$

$$F(X_3 \mid X_4) = \frac{MSR(X_3 \mid X_4)}{MSE(X_3, X_4)}$$

قيمة F الجزئية للمتغير X_1 بوجود X_4

S.V	DF	SS	MS	F
R(X1,X_4)	2	2641.0010		
R(X_4)	1	1831.8962		
R(X_1 |X_4)	1	809.1048	809.1048	**108.22**
ERROR(X_1 ,X_4)	10	74.7621	7.4762	
TOTAL	12	2715.7631		

قيمة F الجزئية للمتغير X_2 بوجود X_4

S.V	DF	SS	MS	F
R(X_2,X_4)	2	1846.8830		
R(X_4)	1	1831.8962		
R(X_2 |X_4)	1	14.9888	19.9888	**0.1725**
ERROR(X_2 ,X_4)	10	868.8801	86.8880	
TOTAL	12	2715.7631		

قيمة F الجزئية للمتغير X_3 بوجود X_4

S.V	DF	SS	MS	F
R(X_3 ,X_4)	2	2590.0251		
R(X_4)	1	1831.8962		
R(X_3 |X_4)	1	708.1289	708.1289	**40.29**
ERROR(X_3 ,X_4)	10	175.7380	17.5738	
TOTAL	12	2715.7631		

ب- نختار المتغير المستقل الذي له أعلى F جزئية والتي هي (108.22) والخاصة بالمتغير X_1 (كميات الومينات الكالسيوم الثلاثي) التي تزيد عن القيمة الجدولية المقابلة لها حيث أن:

$$F=108.22 > f_{0.01,(1,10)} = 3.23$$

لذا فإن X_1 نختارها وندخلها في المعادلة التي تحتوي على X_4 ثم نحسب هذه المعادلة التي كانت:

$$\hat{Y} = 103.097 + 1.940X_1 - 6.614X_4$$

الخطوة الثالثة:

نحسب قيمة F الجزئية لكل من المتغيرات المستقلة الباقية بوجود X_1 , X_4 أي نحسب F $(X_2|X_1, X_4)$ ، $F(X_3|X_1, X_4)$ كما في الجدولين التاليين:

قيمة F الجزئية للمتغير X_2 بوجود X_1 , X_4

S.V	DF	SS	MS	F
$R(X_1,X_2.X_4)$	3	2667.7904		
$R(X_1,X_4)$	2	2641.0110		
$R(X_2 \mid X_1, X_4)$	1	26.77	26.7744	**5.02**
$ERROR(X_1,X_2,X_4)$	9	47.9727	5.3303	
TOTAL	12	2715.7631		

قيمة F الجزئية للمتغير X_3 بوجود X_1 , X_4

S.V	DF	SS	MS	F
$R(X_1,X_3.X_4)$	3	2664.9270		
$R(X_1,X_4)$	2	2641.0110		
$R(X_3 \mid X_1, X_4)$	1	23.9160	23.9160	**4.23**
$ERROR(X_1,X_3,X_4)$	9	50.8361	5.6485	
TOTAL	12	2715.7631		

بما أن المتغير X_2 له أعلى قيمة F جزئية (5.02) حيث:

$$F = 5.02 > f_{0.01,(1,9)} = 3.23$$

لذا نختار X_2 ليضاف للمعادلة السابقة لتصبح لدينا المعادلة الجديدة:

$$\hat{Y} = 71.6480 + 1.4519X_1 + 0.4161X_2 - 0.2365X_4$$

الخطوة الرابعة:

نحسب قيمة F الجزئية للمتغير الباقي X_3 مع وجود X_1 , X_2 $X_{4,}$ في المعادلة أي أننا نحسب

قيمة $F(X_3 | X_1 X_2 X_4)$ وحسب الآتي:

قيمة F الجزئية إلى X_3 بوجود X_1 , X_2 , X_4

S.V	DF	SS	MS	F
$R(X_1, X_2, X_3, X_4)$	4	2667.8995		
$R(X_1,X_2,X_4)$	3	2667.7904		
$R(X_3 \| X_1, X_2, X_4)$	1	0.1091	0.1091	**0.02**
$ERROR(X_1,X_2,X_3.X_4)$	8	47.8637	5.9830	
TOTAL	12	2715.7631		

وبما أن قيمة F الجزئية هنا إلى X_3 صغيرة (0.02) وهي أقل من الجدولية المقابلة لها، أي أن:

$$F = 0.02 < f_{0.01, (1, 8)} = 3.46$$

وهذا يعني أن تأثير المتغير X_3 (الومينات الكالسيوم الرباعي الحديدي) غير معنوي لذا فإنه لا يدخل إلى معادلة الإنحدار.

وبناءً على ذلك، فإن أفضل معادلة إنحدار في هذه الحالة هي:

$$\hat{Y} = 71.6480 + 1.4519X_1 + 0.4161X_2 - 0.2365X_4$$

وبإعتماد هذا النموذج، فإنه بالإمكان إحتساب قيمة معامل التحديد R^2 وفقاً لذلك لتكون

$$R^2 = SSR/SST = 2667.7904/2715.7631 = 0.9823$$

أي أن حوالى 98.23 % من التباينات الموجودة في المتغير المعتمد Y (الحرارة المنبعثة) تعود أسبابها إلى تأثير المتغيرات الثلاثة X_1 (كمية الومينات الكالسيوم الثلاثي) و X_2 (كمية سليكات الكالسيوم الثلاثي) و X_4 (كمية سليكات الكالسيوم الثنائي).

وقد نقوم بتحديد قيمة معامل التحديد المصحح $\overline{R}^2$ وفقاً للصيغة:

$$\begin{aligned}\overline{R}^2 &= 1-\frac{n-1}{n-k}\left(1-R^2\right)\\ &= 1-\frac{12-1}{12-3}\left(1-0.9823\right)\\ &= 0.9784 \qquad (97.84\%)\end{aligned}$$

وقد لا نلاحظ تغيير كبير في القيمتين لكون عدد المتغيرات المستقلة المعتمدة ليس كبير جداً وبعيداً عن الواحد. أي أنه لا فروقات كبيرة ما بين درجة الحرية الكلية (n – 1) ودرجة الحرية للخطأ التجريبي (n–k).

والمثال التالي يتضمن بيانات فعلية لتجربة بمتغيرات مستقلة بعضها ذات قياس كمي والبعض الآخر ذات قياس غير كمي لغرض إستعراض كيفية التعامل مع مثل هذه المتغيرات. والنتائج مأخوذة عن دراسة في جامعة بوسطن الأمريكية[2] عام 2013.

مثال:

في دراسة للعلاقة ما بين ضغط الدم Y كمتغير معتمد ودرجة السمنة Body Mass Index (BMI) كمتغير مستقل X_1 أظهرت وجود إرتباط موجب بمعنوية عالية من خلال نموذج الإنحدار البسيط لبيانات مأخوذة عن عينة بحجم (n = 3,539) حيث كانت النتيجة:

$$\hat{Y} = 108.28 - 0.67\ (BMI)$$

وفي ضوء ذلك تم تطوير الدراسة لنموذج إنحدار متعدد بإعتماد المتغيرات التالية كمتغيرات مستقلة:

BMI = X_1

العمر = X_2

الجنس (ذكر=1 / أنثى= 0) = X_3

تناول أدوية الضغط (نعم =1 / كلا = 0) = X_4

وبذلك يتضح لنا أن المتغيرين X_1 و X_2 بمقياس كمي نسبي فيما كون المتغيران X_3 و X_4 بمقياس إسمي.

والنتيجة كانت:

$$\hat{Y} = 68.15 + 0.58\,X_1 + 0.65\,X_2 + 0.94\,(X_3=1) + 6.44\,(X_4=1).$$

أو يمكننا كتابتها بالشكل التالي:

$$\hat{Y} = 68.15 + 0.58(BMI) + 0.65(Age) + 0.94(Male) + 6.44(Trt)$$

والتي تشير (من خلال إختبار T) إلى معنوية عالية (p-value = 0.0001) لمعاملات الإنحدار بالنسبة إلى جميع المتغيرات عدا متغير الجنس الذكوري X_3 حيث نجد أن (p-value = 0.1133).

وهنا يظهر لنا أن العمر Age هو الأعلى معنوية من بين المتغيرات المستقلة، يتبعه في ذلك BMI ومن ثم المعالجة وأخيراً الجنس الذكوري. كما نلاحظ أن قيمة معامل الإنحدار بالنسبة إلى المتغير BNI قد أصبحت 0.58 مقابل 0.67 أي بإنخفاض قدره 0.13 نتيجة تأثير المتغيرات الثلاثة الأخيرة على قيمة ضغط الدم.

وفيما يخص تفسير النتائج وفقاً لما جاء في أعلاه، فإنها تشير إلى أن زيادة BMI وحدة واحدة يرتبط بزيادة في ضغط الدم بمقدار 0.58 من الوحدات بتثبيت قيم العمر والجنس الذكوري وحالة المعالجة. كذلك يمكننا القول أن الذكور لديهم ضغط دم أعلى نسبياً من الإناث بنحو 0.94 من الوحدات بتثبيت قيمة BMI والعمر وحالة المعالجة.

وبإعتماد معادلة الإنحدار أعلاه نستطيع، بطبيعة الحال، من تقدير قيمة ضغط الدم كدالة لمجموعة المتغيرات المستقلة الأربعة وبالشكل الذي نريده. وعلى سبيل المثال، نجد أن القيمة التقديرية لضغط الدم لدى شخص ذكر عمره 50 سنة ويتسم بمقياس بدانة BMI =25 ولا يخضع لعلاج هي:

$$\hat{Y} = 68.15 + 0.58\ (25) + 0.65\ (50) + 0.94\ (1) + 6.44\ (0) = 116.09$$

بينما هذه القيمة لدى أنثى تخضع للعلاج وبنفس العمر ومقياس البدانة ستكون:

$$\hat{Y} = 68.15 + 0.58\ (25) + 0.65\ (50) + 0.94\ (0) + 6.44\ (1) = 121.59.$$

تحليل المركبات الرئيسية

Principal Components Analysis (PCA)

إن أول من طرح موضوع طريقة تحليل المركبات الرئيسية هو كارل بيرسون Karl Pearson وذلك عام 1901 لأهميتها آنذاك للمختصين في مجال علم الأحياء القياسي Biometrics. أعقبه هوتلنك Hottling عام 1931 بوصف طرق عملية في هذا الجانب.

إن من أهم أسباب تطبيق تحليل المركبات الرئيسية هو لإستخدامها أداةً لفحص البيانات متعددة المتغيرات. ويمكن من خلالها تكوين متغيرات جديدة تسمى علامات المركبات الرئيسية Principal Components Scores حيث يمكن حساب قيمها. إن فحص وضعية الشكل الناتج غالباً ما يعطينا الإنطباع بتحقق حالة اللاطبيعي في البيانات التي نحن بصدد تحليلها.

وفي تحليل المركبات الرئيسية يتم إستخدام اسلوب رياضي يقوم على أساس تحويل مجموعة من المتغيرات التوضيحية المترابطة فيما بينها إلى مجموعة جديدة من المتغيرات غير المترابطة (أو المتعامدة Orthogonal) تدعى المركبات الرئيسية. وكل واحدٍ من هذه المتغيرات الجديدة عبارة عن توليفة رياضية خطية تضم جميع المتغيرات التوضيحية الأصلية ويمكن استخدامها كمدخلات لبرامج الرسومات والأشكال البيانية وطرق تحليل أخرى لمتعدد المتغيرات.

يمكن تنفيذ التحليل هذا باستخدام مصفوفة (التباين-التباين المشترك) Variance-Covariance Matrix أو مصفوفة الإرتباط Correlation Matrix للمتغيرات التوضيحية. وإن نوع المصفوفة المفضل استخدامها يعتمد في الغالب على طبيعة المتغيرات قيد التحليل. فإذا كانت هذه المتغيرات بوحدات متشابهة، يمكن إستخدام مصفوفة (التباين-التباين المشترك). أما إذا كانت الحالة عكس ذلك، فمن الأجدر إعتماد مصفوفة الإرتباط.

بالنسبة للمتغيرات الجديدة (المركبات الرئيسية) يمكننا، أحياناً وليس دائماً، إعطاء تفسير لها. ولذلك لا يمكننا دائماً التوقع بأن نكون قادرين على تفسيرها. بل أنه عندما نعطي تفسيراً لها فإن ذلك يعتبر شيئاً إيحابياً إضافياً وغير متوقع لوظيفتها الرئيسية وهي استخدامها أداةً في متابعة التحليل بطرق أخرى سواءً أمكن تفسيرها أم لا.

ولا يخفى كون تحليل المركبات الرئيسية في العادة مفيداً للباحثين الذين يريدون تقسيم مجموعة الوحدات التجريبية الرئيسية إلى مجاميع فرعية حيث أن الوحدات المتماثلة إلى حدٍ ما تكون ضمن المجموعة الفرعية الواحدة. وفي هذه الحالة، فإن علامات المركبات الرئيسية Principal Components Scores يمكن استخدامها بمثابة مدخل إلى برامج العنقدة Clustering programs مختصرين الوقت والجهد بهذا الإتجاه. والأكثر من ذلك، فإن

علامات المركبات الرئيسية يمكن، بل ويفضل، توظيفها في المساعدة للتحقق من نتائج برامج العنقدة.

طبيعة المركبات الرئيسية :

المركبات الرئيسية عبارة عن توليفات خطية من جميع المتغيرات التوضيحية X_1 , X_2,....., X_p تحددها المتجهات المميزة ($\underline{a_i}$) Eigen Vectors والتي ترتبط بالقيم المميزة (λ_i) Eigen Values (أو الجذور المميزة Characteristic Roots) الناتجة من مصفوفة التباين – التباين المشترك أو مصفوفة الإرتباط. ولابد من أن نذكر هنا بأن عدد هذه المركبات الرئيسية في الأصل هو بعدد المتغيرات المستقلة. وصيغتها الرياضية بشكلٍ عام هي:

$$PC_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_p \qquad , i = 1, 2, ..., p$$

ويمكن التعبيرعن جميع هذه المركبات الرئيسية بصيغة المصفوفات وهي

$$\begin{bmatrix} PC_1 \\ PC_2 \\ \vdots \\ PC_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \mathrm{A'X}$$

حيث أن كل عمود من المصفوفة A يمثل المتجه المميز الذي يقابل القيمة المميزة المرتبط بها والناتج عنها كما سنرى ذلك لاحقاً.

ومن الجدير بالذكر أن المركبة الرئيسية الأولى ستكون عبارة عن توليفة خطية من المتغيرات التوضيحية بمعاملات المتجه المميز المقابل للقيمة المميزة الأولى λ_1 وهي أكبر قيمة مميزة حيث أن هذه القيم تكون عند مقارنتها مع بعضها بالشكل التالي:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots\cdots\cdots \geq \lambda_p$$

ومع أن عدد المركبات الرئيسية، وكما ذكرنا، هي بعدد المتغيرات التوضيحية، إلا أننا لا نستخدم سوى عدد قليل منها سواءً في دالة الإنحدارالمتعدد الخطية أو في طرق أخرى في التحليل متعدد المتغيرات. وهذه الحالة تدخل ضمن عملية تخفيض إتجاهات التحليل الإحصائي للبيانات ومعطياتها Reduction of Dimensionality كما يبدو ذلك واضحاً في بعض طرق التحليل المتعدد وبالأخص في التحليل العاملي. وهذا التخفيض عادةً ما يتم وفقاً لقرار إستبعاد المركبات الرئيسية الضعيفة من حيث تباينها المتمثل بالقيمة المميزة λ وبالأخص من تكون قيمتها أقل من الواحد.

وفيما يتعلق بدالة الإنحدار المتعدد الخطية، فإن الدافع لإستخدامها هو معالجة ظهور حالة التعدد الخطي (الإرتباط المتعدد) Multicollinearity فيما بين المتغيرات التوضيحية وذلك وفقاً للنموذج الآتي:

$$Y = \alpha_0 + \alpha_1 PC_1 + \alpha_2 PC_2 + ... + \alpha_p PC_p$$

وهنا نلاحظ إحلال المركبات الرئيسية محل المتغيرات التوضيحية حيث أنها تتسم بخصوصية التعامدية والتي تعكس عدم وجود أي إرتباط فيما بينها. وهذه السمة ضرورية لأنها تحقق تقديرات صحيحة لمعلمات النموذج.

خطوات الحسابات:

1) نقوم باحتساب مصفوفة التباين – التباين المشترك أو مصفوفة الإرتباط للمتغيرات التوضيحية وذلك طبقاً لكل من الحالتين التاليتين

أ- إذا كانت وحدات قياس المتغيرات التوضيحية متشابهة، فإننا نتجه إلى حساب مصفوفة التباين – التباين المشترك وهي:

$$S = \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1p} \\ V_{21} & V_{22} & \cdots & V_{2p} \\ \vdots & & & \\ V_{p1} & V_{p2} & \cdots & V_{pp} \end{bmatrix}$$

حيث أن:

$$V_{ii} = \frac{S_{ii}}{n-1} \quad , \quad V_{ij} = \frac{S_{ij}}{n-1}$$

$$S_{ii} = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}$$

$$S_{ij} = \sum X_i X_j - \frac{\left(\sum X_i\right)\left(\sum X_j\right)}{n}$$

ب- وإذا كانت وحدات القياس هذه مختلفة، فيستحسن في هذه الحالة تحويل المتغيرات التوضيحية إلى الحالة القياسية بوسط حسابي (0) وتباين (1) وهذا يتطابق تماماً مع إستخدامنا لمصفوفة الإرتباط بدلاً من مصفوفة التباين وهي:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & & & \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

2) إيجاد القيم (الجذور) المميزة λ_i من خلال حل المعادلة المميزة للمصفوفة R وهي:

$$|R-\lambda I| = \begin{vmatrix} 1-\lambda & r_{12} & \cdots & r_{1p} \\ r_{21} & 1-\lambda & \cdots & r_{2p} \\ \vdots & & & \\ r_{p1} & r_{p2} & \cdots & 1-\lambda \end{vmatrix} = 0$$

والتي ستكون:

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots\cdots > \lambda_p$$

3) إيجاد المتجه المميز الأول $\underline{a_1}$ المقابل للقيمة المميزة الأولى λ_1 من المعادلة

$$(R-\lambda_1)\underline{a}_1 = 0$$

ويتم إختيار قيم عناصر هذا المتجه المميز بشرط أن يتحقق لدينا:

$$\underline{a_1}' \, \underline{a_1} = 1$$

وفي هذا السياق، فقد نعتمد المتجه المميز القياسي بمثابة المتجه المميز الأول وهو:

$$\underline{a_1}' = \left[\frac{a_1}{\sqrt{\sum a_i^2}} \quad \frac{a_2}{\sqrt{\sum a_i^2}} \quad \cdots \quad \frac{a_p}{\sqrt{\sum a_i^2}} \right]$$
$$= [a_{11} \quad a_{12} \quad \cdots \quad a_{1p}]$$

وبذلك تكون المركبة الرئيسية الأولى بالصيغة:

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \ldots\ldots + a_{1p}X_p$$

4) إيجاد المتجه المميز الثاني $\underline{a_2}$ المقابل للقيمة المميزة الأولى λ_2 من المعادلة :

$$(R-\lambda_2)\underline{a}_2 = 0$$

وبذلك تكون المركبة الرئيسية الأولى بالصيغة:

$$PC_2 = a_{21}X_1 + a_{22}X_2 + \ldots\ldots + a_{2p}X_p$$

ويتم إختيار قيم عناصر هذا المتجه المميز بشرط أن يتحقق لدينا:

$$\underline{a_2}' \, \underline{a_2} = 1 \qquad , \qquad \underline{a_1}' \, \underline{a_2} = 0$$

والذي هو شرط إعتبار PC_1 و PC_2 متعامدتين.

5) نعيد الإجراءات في الخطوة (4) أعلاه لتحديد المتجه المميز الثالث $\underline{a_3}$ بشرط أن يتحقق لدينا

$$\underline{a_3}' \, \underline{a_3} = 1 \qquad , \qquad \underline{a_1}' \, \underline{a_3} = 0 \quad , \quad \underline{a_2}' \, \underline{a_3} = 0$$

والذي هو شرط إعتبار PC_3 متعامدة مع كلٍ من PC_1 و PC_2

ونستمر هكذا حتى نكمل جميع المركبات الرئيسية واحدةً بعد أخرى وبشرط تحقق حالة تعامد كل واحدة جديدة مع كل ما سبقها.

خواص المركبات الرئيسية:

من المفيد هنا معرفة الجوانب التالية حول بعض خصائص القيم (الجذور) المميزة وهي

1) في حالة إستخدامنا لمصفوفة التباين–التباين المشترك فإن:

$$\sum \lambda_i = Trace(S) = \sum V_{ii}$$

2) في حالة إستخدامنا لمصفوفة الإرتباط فإن:

$$\sum \lambda_i = Trace(R) = p$$

3) نفس الشئ بالنسبة لحاصل ضرب القيم المميزة ليكون مساوياً لمحددة المصفوفة S أي أن:

$$|S| = \Pi_{i=1}^{p} \lambda_i$$

وهي خاصية مهمة جداً في تفسيرات المركبات الرئيسية سيما إذا ما عرفنا أن القيمة المميزة هي بمثابة التباين للمركبة الرئيسية المقابلة لها.

4) إن الأهمية النسبية للمركبة الرئيسية في وصف النموذج تقاس بما يلي

$$\frac{Var(PC_i)}{\sum Var(PC_i)} = \frac{\lambda_i}{\sum \lambda_i}$$

ولذلك لو كانت هناك قيمتين مميزتين أو أكثر متساوية في القيمة، فإنها تكون متساوية في الأهمية النسبية.

5) وحيث أن المركبات الرئيسية متعامدة وبدون أي إرتباط فيما بينها، فإن مصفوفة التباين لها تكون بالصيغة التالية:

$$V(PC) = \begin{bmatrix} V(PC_1) & 0 & \cdots & 0 \\ & V(PC_2) & \cdots & 0 \\ & & & \vdots \\ & & & V(PC_p) \end{bmatrix}$$

$$= \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ & \lambda_2 & \cdots & 0 \\ & & & \vdots \\ & & & \lambda_p \end{bmatrix}$$

وبالتالي من الواضح أن يكون لدينا

$$|V(PC)| = \Pi_{i=1}^{p} \lambda_i$$

بعض إستخدامات المركبات الرئيسية:

1. إن إستخدام المركبات الرئيسية في التحليل المتعدد يساهم في تحديد العوامل الأولية من حيث أهميتها في التحليل إضافة إلى إعطاء فكرة عن مجاميع المتغيرات التي تشكل معاً بعداً خاصاً بها وكما يظهر ذلك في التحليل العاملي حيث أن كل عامل يمكن أن يحمل مسمى يعكس سمة هذه المجاميع.
2. حيث أن المركبات الرئيسية متعامدة فيما بينها جميعاً، فإن إستخدامها في تحليل الإنحدار المتعدد سوف يستبعد حالة التعدد الخطي (الإرتباط المتعدد) وأثره على دقة التقدير مما يؤدي إلى دقة التنبؤ.

عيوب المركبات الرئيسية:

1. ليس للمكونات الرئيسية أي تفسير منطقي واضح في أغلب الأحيان.
2. تعتمد هذه الطريقة على إفتراض التعدد الخطي ما بين المتغيرات التوضيحية. وبذلك فإن إستخدامها في حالة عدم وجود هذه الحالة، سيعطي نتائج تبتعد نسبياً عن الصواب.
3. تعتمد هذه الطريقة على حدس شخصي في بقاء المكونة أو خروجها من التحليل وهو متعلق بقدرالقيم (الجذور) المميزة ولا توجد أية وسيلة يمكن إعتمادها بشكلٍ ثابت في هذا الموضوع.
4. إن إستبعاد بعض المكونات الرئيسية من التحليل يعني إهمال جزء من المعلومات التي قد تكون مفيدة في التحليل النهائي.
5. إن تطبيق هذه الطريقة على القيم المعيارية يعطي نتائج مختلفة عن ما لو تم تطبيقها على المتغيرات الأصلية.
6. الإختبارات المختلفة للمتجهات المميزة (المعدلة) يعطي نوعاً ما أيضاً نتائج مختلفة.

تحليل الإنحدار بالمكونات الرئيسية:

وكما ذكرنا سابقاً، فإن الغرض من إستخدام المكونات الرئيسية في تحليل الإنحدار الخطي المتعدد هو إستبعاد الأثر السلبي للتعدد الخطي في حالة وجوده مع المتغيرات التوضيحية (المستقلة).

ومعادلة الإنحدار تكون وفقاً لذلك:

$$Y = \alpha_0 + \alpha_1 PC_1 + \alpha_2 PC_2 + \dots\dots\dots + \alpha_p PC_p$$

$$= \begin{matrix} \alpha_0 + \alpha_1 \left(a_{11}X_1 + a_{12}X_2 + \dots\dots + a_{1p}X_p\right) \\ + \alpha_2 \left(a_{21}X_1 + a_{22}X_2 + \dots\dots + a_{2p}X_p\right) \\ \vdots \\ + \alpha_p \left(a_{p1}X_1 + a_{p2}X_2 + \dots\dots + a_{pp}X_p\right) \end{matrix}$$

$$
= \begin{aligned}
&\alpha_0 + (\alpha_1 a_{11} + \alpha_2 a_{21} + ... + \alpha_p a_{p1}) X_1 \\
&\quad + (\alpha_1 a_{12} + \alpha_2 a_{22} + ... + \alpha_p a_{p2}) X_2 \\
&\quad \vdots \\
&\quad + (\alpha_1 a_{1p} + \alpha_2 a_{2p} + ... + \alpha_p a_{pp}) X_p
\end{aligned}
$$

بينما معادلة الإنحدار بصيغتها الأصلية تكون:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

لذلك ستكون معلمات الإنحدار الأصلية عند العودة اليها بالتحليل حسب الآتي:

$$
\begin{aligned}
\beta_0 &= \alpha_0 \\
\beta_1 &= \alpha_1 a_{11} + \alpha_2 a_{21} + ... + \alpha_p a_{p1} \\
\beta_2 &= \alpha_1 a_{12} + \alpha_2 a_{22} + ... + \alpha_p a_{p2} \\
&\vdots \\
\beta_p &= \alpha_1 a_{1p} + \alpha_2 a_{2p} + ... + \alpha_p a_{pp}
\end{aligned}
$$

تحقيق التعامدية للمركبات الرئيسية:

إنطلاقاً من وضع المركبات الرئيسية بصيغة المصفوفة:

$$
\begin{bmatrix} PC_1 \\ PC_2 \\ \vdots \\ PC_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = A'X
$$

ونبدأ بالمركبة الرئيسية الأولى:

$$PC_1 = a_{11} X_1 + a_{12} X_2 + ... + a_{1p} X_p = \underline{a_1}' \underline{X}$$

وتباين هذه المركبة يكون:

$$V(PC_1) = V(\underline{a_1}' \underline{X}) = a_1' S a_1$$

وهي ما نريد تعظيمه من خلال إدخال ما يسمى بمضاعف لاكرانج Lagrange Multiplier إلى عملية أخذ المشتقة لهذا التباين مع الأخذ بنظر الإعتبار القيود المصاحبة لهذه المركبة. وفي أدناه توضيحاً لذلك قبل الخوض في العملية المطلوبة للمركبة الرئيسية الأولى.

لو إفترضنا أن المقصود بتعظيمه هي الدالة $f(x)$ ولدينا القيد $g(x)=C$. وبعد إدخال مضاعف لاكرنج سيتم إستحداث دالة جديدة تحتوي جميع هذه الأمور وهي:

$$h(x,\lambda)=f(x)-\lambda[g(x)-C]$$

لاحظ أننا عند إستخدام مضاعف لاكرنج لم نزيد ولم ننقص شيئاً من الدالة $f(x)$ لأننا طرحنا منها ما قيمته صفراً بسبب القيد $g(x)=C$.

ثم نأخذ المشتقة لطرفي المعادلة أعلاه بالنسبة إلى x فيكون لدينا:

$$\frac{\partial h(x,\lambda)}{\partial x}=\frac{\partial f(x)}{\partial x}-\lambda\frac{\partial g(x)}{\partial x}=0$$

ملاحظة:

إذا كان المطلوب تعظيم الدالة $f(x)$ علينا أن نحرص بأن تكون λ هي الأعظم والعكس بالعكس.

وبالنسبة للمركبات الرئيسية، سنقوم بعرض توضيحي لكل مركبة تحديداً:

المركبة الرئيسية الأولى:

تكون لدينا الدالة:

$$\phi_1=\underline{a_1'}S\underline{a_1}-\lambda\left(\underline{a_1'}\underline{a_1}-1\right)$$

ويكون الحل المطلوب لهذه المعادلة بأخذ مشتقتها بالنسبة إلى a_1 ومساواتها للصفر هو الآتي:

$$\frac{\partial\phi_1}{\partial a_1}=2Sa_1-2\lambda a_1=0$$

$$(S-\lambda I)a_1=0$$

ويتم حل هذه المعادلة على أساس أن $a_1\neq 0$ وهذا يعطينا أن:

$$\left|S-\lambda I\right|=0$$

وهذا يعكس لنا بأن λ تمثل القيمة المميزة لمصفوفة التباين S وأن a_1 تمثل المتجه المميز المناظر للقيمة المميزة λ والذي علينا تحديده من خلال العودة إلى النتيجة أعلاه

$$(S-\lambda I)a_1=0$$

$$Sa_1-\lambda Ia_1=0$$
$$Sa_1=\lambda Ia_1$$

وبضرب الطرفين من اليسار بالقيمة a_1' يكون لدينا:

$$a_1'Sa_1=a_1'\lambda Ia_1=\lambda Ia_1'a_1=\lambda a_1'a_1=\lambda=\lambda_1$$

ويتم هنا حل هذه المعادلة بالنسبة إلى a_1 والتي تناظر القيمة المميزة العظمى λ_1.

ملاحظة:

لقد إستخلصنا من أعلاه بأن تباين المركبة الرئيسية الأولى هو $a_1'Sa_1=\lambda_1$ وهو الأعظم ومتوازناً مع الشرط $a_1'a_1=1$ وبعدها ننتقل للمركبة الرئيسية الثانية والتي هي:

$$PC_2=a_{21}X_1+a_{22}X_2+.......+a_{2p}X_p=a_2'X$$

وتباين هذه المركبة يكون:

$$V(PC_2)=V(a_2'X)=a_2'Sa_2$$

وهي ما نريد تعظيمه من خلال إدخال مضاعف لاكرانج Lagrange Multiplier إلى عملية أخذ المشتقة لهذا التباين مع الأخذ بنظر الإعتبار القيود المصاحبة لهذه المركبة. وهي

$$a_2'a_2=1$$

مع تثبيت شرط التعامدية مع المركبة الرئيسية الأولى وهو $a_2'a_1=0$.

المركبة الرئيسية الثانية:

تكون لدينا الدالة:

$$\phi_2=a_2'Sa_2-\lambda(a_2'a_2-1)-2\mu\left(a_2'a_1-0\right)$$

ويكون الحل المطلوب لهذه المعادلة بأخذ مشتقتها بالنسبة إلى a_2 ومساواتها للصفر هو الآتي:

$$\frac{\partial\phi_2}{\partial a_2}=2Sa_2-2\lambda a_2-2\mu a_1=0$$

$$(S-\lambda I)a_2=\mu a_1$$

ويتم حل هذه المعادلة على أساس $a_2 \neq 0$:

$$Sa_2 - \lambda I a_2 = \mu a_1$$

وبضرب الطرفين من اليسار بالقيمة a_1' يكون لدينا:

$$Sa_1'a_2 - \lambda a_1'a_2 =_1 \mu a_1' a_1$$
$$0-0=\mu$$
$$\mu = 0$$

وبذلك يكون لدينا:

$$(S-\lambda I)a_2 = 0$$

وبما أن $a_2 \neq 0$ سيكون لدينا:

$$|S-\lambda I| = 0$$

وبضرب طرفي المعادلة $(S-\lambda I)a_2 = 0$ بالقيمة a_1' يكون لدينا:

$$a_2'Sa_2 - a_2'\lambda I a_2 = 0$$

$$a_2'Sa_2 - \lambda a_2'a_2 = 0$$
$$a_2'Sa_2 = \lambda = \lambda_2$$

ويتم هنا حل هذه المعادلة بالنسبة إلى a_2 والتي تناظر القيمة المميزة العظمى الثانية λ_2.

وبعدها ننتقل للمركبة الرئيسية الثالثة والتي هي:

$$PC_3 = a_{31}X_1 + a_{32}X_2 + ... + a_{3p}X_p = a_3'X$$

وتباين هذه المركبة يكون:

$$V(PC_3) = V(a_3'X) = a_3' S a_3$$

وهي ما نريد تعظيمه من خلال إدخال مضاعف لاكرانج Lagrange Multiplier إلى عملية أخذ المشتقة لهذا التباين مع الأخذ بنظر الإعتبار القيود المصاحبة لهذه المركبة. وهي:

$a_3' a_3 = 1$ مع تثبيت شرط التعامدية مع المركبة الرئيسية الأولى وهو $a_3' a_1 = 0$ ومع المركبة الرئيسية الثانية $a_3' a_2 = 0$.

المركبة الرئيسية الثالثة:

تكون لدينا الدالة:

$$\phi_3 = a_3' S a_3 - \lambda(a_3' a_3 - 1) - 2\mu_1\left(a_3' \, a_1 - 0\right) - 2\mu_2\left(a_3' \, a_2 - 0\right)$$

ويكون الحل المطلوب لهذه المعادلة بأخذ مشتقتها بالنسبة إلى a_3 ومساواتها للصفر هو الآتي:

$$\frac{\partial \phi_3}{\partial a_3} = 2Sa_3 - 2\lambda a_3 - 2\mu_1 a_1 - 2\mu_2 a_2 = 0$$

$$(S - \lambda I)a_3 - \mu_1 a_1 - \mu_2 a_2 = 0$$

ويتم حل هذه المعادلة على أساس $a_3 \neq 0$.

وبضرب الطرفين من اليسار بالقيمة a_1' يكون لدينا:

$$a_1'(S - \lambda I)a_3 - a_1'\mu_1 a_1 - a_1'\mu_2 a_2 = 0$$
$$(S - \lambda I)a_1' a_3 - \mu_1 a_1' a_1 - \mu_2 a_1' a_2 = 0$$

وهذا يعطينا $\mu_1 = 0$

ولو ضربنا نفس الطرفين من اليسار بالقيمة a_2' يكون لدينا:

$$a_2'(S - \lambda I)a_3 - a_2'\mu_1 a_1 - a_2'\mu_2 a_2 = 0$$
$$(S - \lambda I)a_2' a_3 - \mu_1 a_2' a_1 - \mu_2 a_2' a_2 = 0$$

وهذا يعطينا $\mu_2 = 0$ وبذلك يكون لدينا:

$$(S - \lambda I)a_3 = 0$$

وبما أن $a_3 \neq 0$ سيكون لدينا:

$$|S - \lambda I| = 0$$

وبضرب طرفي المعادلة $(S - \lambda I)a_3 = 0$ من اليسار بالقيمة a_3' يكون لدينا:

$$a_3' S a_3 - a_3' \lambda I a_3 = 0$$
$$a_3' S a_3 - \lambda a_3' a_3 = 0$$
$$a_3' S a_3 = \lambda = \lambda_3$$

وهكذا لبقية المركبات الرئيسية حيث نستمر بنفس النهج التصاعدي لمضاعف لاكرنج.

مثال 1 (حالة متغيرين)

البيانات التالية تمثل الدخل الشهري X_1 بعملة معينة و X_2 تمثل الخدمة بالسنوات و Y تمثل الزيادة التشجيعية لعينة من خمسة عاملين. ومقياس جميع المتغيرات هنا هو كمي (نسبي أو فئوي).

العامل	Y	X_1	X_2
1	8	102	4
2	9	104	5
3	7	101	7
4	9	93	1
5	10	100	3

سنستخدم هنا مصفوفة الإرتباط وهي:

$$R = \begin{bmatrix} 1 & 0.7483 \\ 0.7483 & 1 \end{bmatrix}$$

$$|R - \lambda I| = \begin{vmatrix} 1-\lambda & 0.7483 \\ 0.7483 & 1-\lambda \end{vmatrix} = 0$$

$$(1-\lambda)^2 - 0.56 = 0$$
$$\lambda_1 = 1.7483$$
$$\lambda_2 = 0.2517$$

وعلينا تحديد قيم عناصر المتجه $\underline{a_1}$ من خلال الآتي:

$$(R - \lambda I)\,\underline{a_1} = 0$$
$$(R - 1.7483\,I)\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0$$
$$\begin{pmatrix} 1-1.7483 & 0.7483 \\ 0.7483 & 1-1.7483 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0$$
$$\begin{pmatrix} -0.7483 & 0.7483 \\ 0.7483 & -0.7483 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0$$

ومن الواضح هنا أن $a_1 = a_2$ ولذلك لو إفترضنا القيمة (1) لأحداهما ستكون الأخرى مساوية لها وبنفس القيمة. أي أنه يصبح لدينا:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

وهذه القيم الأولية يمكن إستخدامها لتحديد قيم عناصر المتجه المميز الذي يتسم بشرط مجموع مربع قيم العناصر مساوياً إلى 1. ولتحقيق ذلك، نتبع التحويل التالي:

$$\underline{a_1'} = \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} \dfrac{a_1}{\sqrt{\sum a_i^2}} \\ \dfrac{a_2}{\sqrt{\sum a_i^2}} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sqrt{2}} \\ \dfrac{1}{\sqrt{2}} \end{bmatrix}$$

وهي قيم عناصر المتجه القياسي المعتمد الأول.

والآن علينا تحديد قيم عناصر المتجه القياسي الثاني $\underline{a_2}$ من خلال إتباع نفس الإسلوب ولكن بإستخدام القيمة المميزة الثانية وحسب الآتي:

$$(R - \lambda I)\,\underline{a_2} = 0$$

$$(R - 1.7483I)\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0$$

$$\begin{pmatrix} 1-0.2517 & 0.7483 \\ 0.7483 & 1-0.2517 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0$$

$$\begin{pmatrix} 0.7483 & 0.7483 \\ 0.7483 & 0.7483 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0$$

ومن الواضح هنا أن $a_1 = -a_2$ ولذلك لو افترضنا القيمة (1) لأحداهما ستكون الأخرى هي:

(1) أي أنه يصبح لدينا:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

وهذه القيم الأولية يمكن إستخدامها لتحديد قيم عناصر المتجه المميز الذي يتسم بشرطي مجموع مربع قيم العناصر مساوياً إلى 1 وحاصل ضرب المتجهين يساوي 0.0.

ولتحقيق ذلك، نتبع التحويل التالي:

$$\underline{a_1'} = \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} \dfrac{a_1}{\sqrt{\sum a_i^2}} \\ \dfrac{a_2}{\sqrt{\sum a_i^2}} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sqrt{2}} \\ -\dfrac{1}{\sqrt{2}} \end{bmatrix}$$

وهي قيم عناصر المتجه القياسي المعتمد الثاني.

وحيث أن $\frac{1}{\sqrt{2}} = 0.707$ ، فإن المركبتين الرئيسيتين ستكونا بالشكل التالي:

$$PC_1 = 0.707X_1 + 0.707X_2$$
$$PC_2 = 0.707X_1 - 0.707X_2$$

ومن المناسب هنا تجميع جميع القيم المرتبطة بالمركبات الرئيسية التي تم تحديدها وهو ما يتعلق بالمتجهات والقيم المميزة مع الأهمية النسبية لكل مركبة. كل ذلك سيكون ضمن الجدول التالي:

	PC_1	PC_2
X_1	0.707	0.707
X_2	0.707	-0.707
λ	1.7483	0.2517
الأهمية النسبية	87.4%	12.6%

وهنا لو قررنا الإبقاء على المركبات الرئيسية المهمة فقط، فإننا سنكتفي بالمركبة الأولى لكون القيمة المميزة تزيد عن الواحد. وهذا الإجراء هو ما نسميه بتقليص الإتجاهات Reduction of Dimensionality في التحليل.

ولغرض إعطاء توضيح أكثر لهذا الموضوع، سنأخذ المثال التالي لبيانات حقيقية:

مثال 2:

في دراسة G. R. Bryce في جامعة Brigham Young حول إحتمالية وجود علاقة ما بين تصميم خوذة الرأس للاعبي كرة القدم الأمريكية وجروح الرقبة، تم تسجيل 6 قياسات (متغيرات مستقلة) لثلاثة مجاميع من اللاعبين وبواقع 30 لاعب لكل مجموعة وهي كما يلي[(3)]:

المجموعة (1) : تمثل لاعبي المرحلة الثانوية لكرة القدم

المجموعة (2) : تمثل لاعبي المرحلة الجامعية لكرة القدم

المجموعة (3) : تمثل من هم بالعمر الجامعي وليسوا من لاعبي كرة القدم

أما القياسات (المتغيرات المستقلة) الست لاعبين فهي كما يلي:

X_1 = عرض الرأس عند أكبر عمق

X_2 = محيط الرأس

X_3 = المسافة ما بين مقدمة الرأس ومؤخرته عند مستوى العين

X_4 = المسافة ما بين العين وقمة الرأس

X_5 = المسافة ما بين الأذن وقمة الرأس

X_6 = عرض الفك

وجميع هذه المتغيرات بمقياس كمي نسبي وبنفس الوحدات.

تنويه: في هذا المثال تم إعتماد المجموعتين الثانية والثالثة فقط لوجود التجانس العمري بينهما.

ونبدأ العمل بطبيعة الحال بمصفوفة التباين والتباين المشترك Variance – Covariance Matrix للمتغيرات المستقلة الست وهي:

$$S = \begin{bmatrix} 0.320 & 0.602 & 0.149 & 0.044 & 0.107 & 0.209 \\ & 2.629 & 0.801 & 0.666 & 0.103 & 0.377 \\ & & 0.458 & 011 & -0.013 & 0.120 \\ & & & 1.474 & 0.252 & -0.054 \\ & & & & 0.488 & -0.036 \\ & & & & & 0.324 \end{bmatrix}$$

ومجموع العناصر القطرية لهذه المجموعة يمثل مجموع التباين والذي يمثل بدوره مجموع القيم المميزة لهذه المصفوفة. أي أن:

$$\sum_{j=1}^{6} S_{jj} = \sum_{i=1}^{6} \lambda_{ii} = 0.320 + 2.629 + 0.458 + 1.474 + 0.488 + 0.324 = 5.743$$

ومن خلال التحليل تم تحديد القيم المميزة Eigen values الست مع المتجهات المميزة Eigen vectors للقيم المعتمدة وهي الأولى والثانية فقط كونهما توضحان نسبة 81.8% من التباين وكلاهما أكبر من الواحد، وكانت كما يلي:

Eigen value القيمة المميزة	Proportion of Variance نسبة التباين	Cumulative proportion النسبة المتجمعة	Eigen Vectors المتجهات المميزة		
			المتغير	**a_1**	**a_2**
3.323	0.579	0.579	X_1	0.207	-0.142
1.374	0.239	0.818	X_2	0.873	-0.219
0.476	0.083	0.901	X_3	0.261	-0.231
0.325	0.057	0.957	X_4	0.326	0.891
0.157	0.027	0.985	X_5	0.066	0.222
0.088	0.015	1.000	X_6	0.128	-0.187

ووفقاً لذلك فإننا سنكتفي بالمركبة الرئيسية الأولى PC_1 والثانية PC_2 وتكون بالصيغة التالية:

$$PC_1 = 0.207X_1 + 0.873X_2 + 0.261X_3 + 0.362X_4 + 0.066X_5 + 0.128X_6$$

$$PC_2 = -0.142X_1 - 0.219X_2 - 0.231X_3 + 0.891X_4 + 0.222X_5 - 0.187X_6$$

ومن الملاحظ هنا أن X_2 (محيط الرأس) له تأثير واضح ضمن المركبة الرئيسية الأولى بمعامل (0.873) ومثل ذلك بالنسبة إلى X_4 (المسافة ما بين العين وقمة الرأس) ضمن المركبة الرئيسية الثانية بمعامل (0.891). وهذا متوقع حدوثه لكون هذين المتغيرين لهما أكبر تباين (2.629 بالنسبة إلى X_2 و 1.474 بالنسبة إلى X_4) بالمقارنة مع تباينات بقية المتغيرات.

وجدير بالذكر هنا أنه لو كانت هذه المتغيرات بتباينات متقاربة، لكنا قد لاحظنا تقارب كبير في معاملاتها ضمن أي من المركبتين. ومن جهة أخرى، لو كانت تباينات هذين المتغيرين X_2 و X_4 كبيرة جداً نسبياً، لكنا نلاحظ أن المركبة الرئيسية الأولى PC_1 تساوي إلى حدٍ ما المتغير X_2، كما أن المركبة الرئيسية الأولى PC_2 تساوي إلى حدٍ ما المتغير X_4.

نموذج الإنحدار اللوجستي
Logistic Regression Model (LRM)

مقدمة

يستخدم الإنحدار اللوجستي في الغالب لنمذجة إحتمال عائدية وحدة تجربة لمجموعة معينة استناداً إلى معلومة مأخوذة من تلك الوحدة. مثل هذه النماذج يمكن استخدامها لأغراض التمييز. وفي حالة البطاقة الإئتمانية، يمكننا نمذجة إحتمال كون شخصٍ ما ذو خصائص ديموغرافية معينة يقع ضمن مجموعة الجيدين من ناحية خطورة الإئتمان. وبعد تطوير هذا النموذج، بالإمكان استخدامه للتنبؤ بالمجموعة التي ينتمي اليها شخصاً جديداً وفقاً لسماته الديموغرافية المحددة. والشخص الذي يعطي عنه النموذج إحتمالاً يزيد عن 0.5 يحدد بأنه ينتمي إلى مجموعة الجيدين بالنسبة لخطورة الإئتمان.

وطالما أنه نموذج انحدار، فهذا يعني أن هناك متغيرات توضيحية تقابل متغير معتمد لكن هذا المتغير المعتمد قد يكون بقياس رقمي عادي Numerical أو مقياس فئوي إسمي Nominal. ولأننا نتحدث عن إحتمال، فإن هذا هو إحتمال انتساب المتغير المعتمد لفئة معينة. ولذلك، فإنه في حالة كون القياس رقمي فإننا نلجأ إلى تقسيم القيم بحدود مناسبة لغرض تحديد الفئات بموجبها.

إن نماذج الإنحدار اللوجستي تعتبر إسلوباً جيداً للوقوف على كيفية تأثير عددٍ من العوامل على ظهور مشاهدة ثنائية القياس Binary كمتغير معتمد (إستجابة Response). ونقصد بالمتغير الثنائي Binary هو أن هذا المتغير يأخذ قيمتين محتملتين. والأمثلة على ذلك الوفيات (حي/متوفي) والحالة المرضية (سليم/ مريض) والإجابة (نعم/ كلا) والجنس (ذكر/ أنثى) وهكذا.

وفي بعض الأحيان يمكن أن يحصل لدينا متغير تابع (معتمد) مستمر القياس وليس ثنائياً مثل مدة المكوث في المستشفى جراء الولادة. وفي هذه الحالة فقد يتم تقسيم القيم إلى أصغر أو يساوي 48 ساعة مقابل أكثر من 48 ساعة ليصبح القياس عندئذٍ ثنائياً.

وبالنسبة للمشاهدات الثنائية، فقد نجد إسلوب الترميز الرقمي مناسباً بإستخدام القيمتين (صفر/1) لكل مشاهدة ويفضل إستخدام العدد (1) ليعكس وجود الظرف(الحالة) والعدد (صفر) لغيابها. وكمثال على ذلك، ينسب الرقم (1) للمريض والرقم (صفر) لغير المريض (السليم).

ولتوضيح أهمية إستخدام نموذج الإنحدار اللوجستي، إفترض أنه لدينا مشاهدة ثنائية Y (المرض : نعم / كلا) بصفة الإستجابة ومتغير مؤثر آخر X (التعرض : نعم / كلا). وهنا يمكن تسمية X بأنه عامل الخطورة وتسمية Y بأنه عامل الإستجابة تجاه هذه الخطورة.

وفي مثل هذه الحالة، يمكننا إستخدام جدول تقاطعي (2x2) لتقدير الخطورة النسبية Relative Risk وعلى النحو التالي:

		Y	
		Yes (1)	No (0)
X	Yes(1)	a	b
	No(0)	C	d

وحيث أن a , b , c , d تمثل تكرارات لهذه التقاطعات وأن حجم العينة n تكون:

$$n = a + b + c + d$$

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

في حالة الدراسات الفوجية Cohort أو ما نسميه الدراسات التتبعية Prospective.

كذلك يمكن استخدامها لتقدير نسبة الأرجحية Odds Ratio (OR) في حالتي التتبعية أو التقاطعية Case – Control. وتحسب نسبة الأرجحية هذه بأنها

$$OR = \frac{(a)(d)}{(b)(c)}$$

والآن لنفترض أن Y متغير ثنائي بينما X متغير مستمر. وفي هذه الحالة، لا يمكن إستخدام الجدول التقاطعي (2x2).

وفي مثل هذه المسألة، نستخدم الإنحدار اللوجستي البسيط في حالة متغير مؤثر (مستقل) واحد X أو الإنحدار اللوجستي المتعدد في حالة وجود أكثر من متغير مؤثر(مستقل) واحد مثل $X_1 , X_2 , \dots , X_p$ كمتغيرات مستقلة.

والآن دعنا نتناول نموذج الإنحدار اللوجستي البسيط والذي يتضمن متغير مستقل واحد X.

ونموذج الإنحدار اللوجستي يكون مناسباً لدالة الإستجابة Y مقابل X (أو متعدد X_1 , X_2 ,...) مع منحنى يعرف بشكل S أي S – Shaped Curve ويعرف بالدالة اللوجستية Logistic Function ، ويعبر عنها بالآتي:

$$E(Y|X) = \exp(\beta_0 + \beta_1 X)/ [1+ \exp(\beta_0 + \beta_1 X)]$$

حيث أن E(Y|X) هي القيمة المتوقعة إلى Y عند وجود X. وأن " exp " تعبر عن الأساس إلى اللوغاريتم الطبيعي ($\log_e = \ln$) وأن (e = exp = 2.71828). كما أن β_0 و β_1 عبارة عن معامل الإنحدار وهما معلمتان يتم العمل على تقديرهما في النموذج.

ولأن معدل دالة الإستجابة p = E(Y|X) عندما يكون Y متغيراً ثنائياً Binary حينما نستخدم القيمة (صفر) للتعبير عن الفشل والقيمة (واحد) للتعبير عن النجاح والذي تكون p عبارة عن إحتمال ظهوره. ولذلك فإننا نستخدم في العادة صيغة النموذج التالية :

$$P = \exp(\beta_0 + \beta_1 X)/ (1+ \exp(\beta_0 + \beta_1 X)$$

ولكي نحصل على نموذج خطي بالنسبة إلى المعلمات β_0 و β_1 فإننا نحتاج لإستخدام التحويل اللوجستي هنا ويسمى Logistic transformation حيث أن:

$$\text{Logit} = \log_e \left(\frac{p}{1-p}\right) = \ln \left(\frac{p}{1-p}\right)$$

وعلى سبيل المثال، لو كان إحتمال ظهور النجاح لمشاهدة ما هو p = 0.1 فإن:

$$\text{Logit} = \ln \left(\frac{0.1}{0.9}\right) = - 2.1972$$

وفيما يلي توضيح لكيفية الحصول على نموذج خطي في β_0 و β_1. حيث لدينا:

$$P = \exp(\beta_0 + \beta_1 X)/ (1+ \exp(\beta_0 + \beta_1 X)$$

$$1 - p = 1 - \exp(\beta_0 + \beta_1 X)/ (1+ \exp(\beta_0 + \beta_1 X)$$

$$= 1/ (1+ \exp(\beta_0 + \beta_1 X)$$

وبالتالي فإن:

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 X)$$

$$\ln \left(\frac{p}{1-p}\right) = \ln[\exp(\beta_0 + \beta_1 X)] = \beta_0 + \beta_1 X \qquad ....(1)$$

وجدير بالذكر هنا أننا نطلق تعبير "نسبة الأرجحية" Odd's Ratio على المقدار $\frac{p}{1-p}$

أي أن: $OR = \frac{p}{1-p}$

وجدير بالملاحظة أن $logit = \log(OR)$ هو بمثابة المتغير المعتمد في النموذج الخطي كما في المعادلة (1) أعلاه.

ملاحظة:

أينما يذكر اللوغاريتم log فإننا نعني اللوغاريتم الطبيعي ln.

ولو إفترضنا أن المتغير X ثنائي القياس Binary بحيث يكون X=0 عند تواجده و X=1 عند عدم تواجده لتوصلنا إلى حقيقة كون معدل التغير في نسبة الأرجحية OR تستند على تقدير المعلمة β_1 أي b_1.

وفيما يلي توضيح لكيفية حدوث ذلك لغرض التثبيت لهذه الحقيقة. حيث أن:

$$
\begin{aligned}
\text{Log (OR)} &= \log[(\text{odds}|X=1)/(\text{odds}|X=0)] \\
&= \log(\text{odds}|X=1) - \log(\text{odds}|X=0) \\
&= \text{logit}(X=1) - \text{logit}(X=0) \\
&= (\beta_0 + \beta_1 X) - (\beta_0) \quad , \ X=1 \\
&= \beta_1
\end{aligned}
$$

وبذلك فإن المماس β_1 هو عبارة عن log(OR) ولذلك فإن:

$$OR = \exp(\beta_1)$$

في مثل هذه الحالات (أي حالة كون X متغير ثنائي القياس Binary).

والآن يتم تركيزنا على أهمية نموذج الإنحدار اللوجستي من أنه يفحص العلاقة ما بين متغير مستقل واحد أو أكثر و log (OR) المتغيرالمعتمد (الإستجابة) ثنائي القياس.

وإذا ما نظرنا إلى log (OR) يترآى لنا أننا نتبع طريقة معقدة لتفسير البيانات، ولكن هذا يؤدي إلى أسهل إسلوب لتفسير البيانات والتي تتماشى مع قواعد الإحتمالات.

ولتوضيح هذا الجانب، دعنا نفترض البيانات التالية:

GA	28	29	30	31	32	33	34
Prob(BF)	0.60	0.62	0.64	0.66	0.68	0.70	0.72

ملاحظة:

GA يمثل فترة الحمل بالأسابيع وهو المتغير المستقل وذو قياس نسبي كونه كمي متصل

BF يمثل رضاعة الطفل من صدر أمه بعد مغادرته المستشفى وهو متغير معتمد.

وثنائي القياس (yes =1 , No = 0) ومقياسه إسمي.

وعند إستخدامنا لنموذج خطي للربط ما بين المتغيرين، فقد نقول أن:

$$Prob(BF) = \beta_0 + \beta_1 (GA)$$

أي أنه من الواضح كون Prob(BF) (وهي إحتمال رضاعة الطفل من صدر أمه) تزداد وفق دالة خطية بالنسبة إلى GA (مدة الحمل بالأسابيع). وعند تحليل هذه البيانات وفقاً لذلك، نجد أن:

$$Prob(BF) = 0.04 + 0.02 (GA)$$

وعلى هذا الأساس، فإن رضيعاً مع فترة حمله 30 إسبوعاً سيرضع من صدر أمه بعد مغادرتها المستشفى بإحتمال 0.64 وفقاً لهذا النموذج التجميعي.

ملاحظة مهمة (للنموذج التجميعي) Additive Model

نلاحظ هنا أننا إستخدمنا نموذج تجميعي في الإحتمالات. وهذا قد يقودنا إلى نتائج غير منطقية في بعض الأحيان وغير محسوبة خاصة إذا ما إقترب الإحتمال إلى % 0.0 أو إلى 100% لأننا نتوقع أن نجتاز هذين الحدين فيكون لدينا إحتمال أكثر من 100% أو أقل من الصفر (إحتمال سالب) وكلاهما غير منطقي بالنسبة للإحتمالات. دعنا الآن نجري بعض التغيير في نموذجنا الخطي للآتي:

$$Prob (BF) = 0.04 + 0.03 (GA)$$

وهذا سوف يعطينا النتائج التقديرية التالية:

GA	28	29	30	31	32	33	34
Prob(BF)	0.88	0.91	0.94	0.97	1.0	1.03	1.06

وبطبيعة الحال، سنجد صعوبة في إعطاء تفسير عما يعنيه الإحتمال 1.06 وهذا هو السبب الرئيسي الذي يدفعنا لتجنب إستخدام النموذج التجميعي.

وبشكلٍ عام، فإنه ينصح بمحاولة تجنب النموذج التجميعي هذا ما لم يكن هنالك سبب قوي (قناعة) لتوقع كون جميع قيم الإحتمالات المتوقعة ستكون ضمن المدى (80% - 20%).

ونعني بتجنب النموذج التجميعي هو تبني النموذج الضربي للإحتمالات وهو التالي.

النموذج الضربي للإحتمالات A Multiplicative Model

قد يكون من الأجدر إعتماد النموذج الضربي في حالة الإحتمالات مع كونه قد يعاني أيضاً من نفس المشاكل كما هي الحال في النموذج التجميعي ولكن بشكلٍ أقل. وهنا تعتمد عملية الضرب بدلاً من عملية الجمع عند تغيير قيمة الإحتمال. وفيما يلي مثالاً على ذلك.

لنفترض النتيجة التقديرية التالية:

GA	28	29	30	31	32	33	34
Prob(BF)	0.01%	0.03%	0.09%	0.27%	0.81%	2.43%	7.29%

وفي هذا المثال، فإن كل إسبوع إضافي في مدة الحمل يؤدي إلى مضاعفة إحتمال الرضاعة الطبيعية Prob(BF) ثلاث مرات. كما نلاحظ هنا أن النموذج الضربي لا يؤدي إلى إحتمال أقل من الصفر ولكنه قد يؤدي إلى إحتمال أكثر من 1.0. ولغرض تطبيقه علينا أن تكون لدينا القناعة القوية بأن توقع جميع الإحتمالات ستكون بقيم صغيرة، ولتكن أقل من 0.20.

العلاقة بين الأرجحية Odds والإحتمال .Prob

لنفترض حالة الربح والخسارة لفريق كرة قدم. فإذا كانت الأرجحية 3 إلى 1 بجانب فريقك فهذا يعني أنك تتوقع الربح (الفوز) أن يحصل 3 مرات بقدر عدد الخسارة. سنتعامل مع الحالة هذه إضافة إلى الأرجحية 4 إلى 1 ونرى كيف يمكننا تحويل الأرجحية إلى إحتمال وبالعكس. ففي الحالة الأولى يكون الربح ثلاثة مرات مع كل أربعة أشواط أي أن إحتمال الفوز هو 0.75. أما إذا كان إحتمال الفوز هو 0.20 فهذا يعني أنك تتوقع مرة واحدة للفوز مقابل أربع مرات خسارة مع كل خمسة أشواط لعب.

وهنا نستطيع وضع الصيغة التالية لهذه العلاقة ما بين نسبة الأرجحية OR والإحتمال Prob

$$OR = Prob/(1 - Prob) = P/(1 - P)$$

$$Prob = OR/(1 + OR)$$

والآن دعنا نرى تطبيق نموذج الإنحدار اللوجستي البسيط بمتغير مستقل واحد ومن خلال البيانات التالية:

البيانات الأولية		القيم المتوقعة		
GA = X	Prob (BF) = Y	Log (Odds)	Odds(BF)	Prob(BF)
28	2/6= 0.333	- 0.57	0.57	0.362
29	2/5= 0.400	0.01	1.01	0.503
30	7/9 =0.778	0.59	1.80	0.643
31	7/9 =0.778	1.16	3.20	0.762
32	16/20=0.80	1.74	5.70	0.851
33	14/15=0.933	2.32	10.15	0.910

وفي حالة تطبيق نموذج الإنحدار اللوجستي البسيط يكون لدينا:

$P = Y = \exp(\beta_0 + \beta_1 X)/[1 + \exp(\beta_0 + \beta_1 X)]$

$\text{Logit}(P) = \log [P/(1 - P)] = \beta_0 + \beta_1 X$

وبالتالي سنحصل على النموذج المتوقع:

$\text{Logit}(P) = \log [P/(1 - P)] = -16.72 + 0.577 X$

وفي حالة كون $X = GA = 30$ ، فإن هذا النموذج يعطينا:

$\text{Logit}(P) = \log [P/(1 - P) = \log(OR)$

$= -16.72 + 0.577(30) = 0.59$

ولتحويل هذه القيمة إلى Odds يكون لدينا:

$OR = \exp(o.59) = 1.80$

وأخيراً تحويل ذلك إلى الإحتمال (رجوعاً) يكون لدينا:

$\text{Prob.} = 1.80/(1 + 1.80) = 0.643$

وهذه القيمة المتوقعة للإحتمال تعتبر قريبة بشكلٍ معقول من الإحتمال الحقيقي وهو (0.778).

ومن المفيد أيضاً القاء نظرة على نسبة الأرجحية المتوقعة OR (BF) حيث نلاحظ من الجدول بأن نسبة أي OR لصفين متتاليين هي 1.78.

ومثالاً على ذلك نجد أن $3.20/1.80 = 1.78$

وكذلك $5.70/3.20 = 1.78$

وهذه النتيجة ليست من قبيل المصادفة وإنما هي واقع قيمة المقدار

$$\exp(\beta_1) = \exp(0.577) = 1.78$$

وهذه هي الصفة العامة لنموذج الإنحدار اللوجستي. فقيمة الميل في نموذج الإنحدار اللوجستي يمثل log(OR) أو لوغارتم نسبة الأرجحية والتي تمثل (الزيادة/النقصان) في الخطورة عندما تزداد قيمة المتغير المستقل وحدة واحدة.

نموذج الإنحدار اللوجستي المتعدد Multiple Logistic Regression Model

في هذا النموذج يكون لدينا متغيرين مستقلين أو أكثر $X_1 , X_2 , ... , X_k$. وبالتالي، يكون النموذج وفقاً للمعادلة:

$$Y = P = \exp(\beta_0 + \sum \beta_i X_i)/[1 + \exp(\beta_0 + \sum \beta_i X_i)]$$

$$\text{Logit}(P) = \log [P/(1 - P)] = \beta_0 + \sum \beta_i X_i$$

ومن المزايا الجيدة لهذا النموذج أنه يسمح بوجود المتغيرات المستقلة (التوضيحية) بقياسات مختلفة بالنسبة إلى $X_1 , X_2 , ... , X_k$. فقد يكون لدينا متغير أو أكثر فئوي القياس (سواءً إسمي أو رتبي أو فتري). وسنرى ذلك من خلال المثال التالي والذي يتضمن دراسة لتحديد مدى تأثير عوامل مختلفة عديدة ومتنوعة القياس تجاه حالة البدانة لدى المرأة الأردنية.

المثال :

في دراسة لتحديد مدى تأثير عدد من العوامل تجاه حالة البدانة لدى المرأة الأردنية[4]، تم إعتبار مؤشر البدانة والمتمثل بالمقياس BMI بمثابة المتغير المعتمد Y حيث:

BMI يرمز إلى Body Mass Index ويتم حسابه وفقاً للآتي

$$BMI = \text{Body weight (Kg)} / [\text{Body High (meter)}]^2 = Kg/m^2$$

وتعتبر المرأة بدينة في حالة كون المقياس $BMI \geq 30 \ Kg/m^2$ مع مجموعة العوامل المؤثرة X والتي هي بمثابة المتغيرات المستقلة وتشمل:

AGE : ويمثل عمر المرأة بالسنوات ومقياسه كمي نسبي.

ED : ويمثل مستوى التعليم (أساسي أو أقل / ثانوي ED1 /عالي ED2) ومقياسه كمي فئوي ثلاثي.

PAR : ويمثل حجم الولادات (حد أعلى 1 /(3-2) PAR1 / (4 ≤) PAR2) ومقياسه كمي فئوي ثلاثي.

REG : ويمثل المنطقة الجغرافية (المركز / شمال REG1 / جنوب REG2) ومقياسه إسمي ثلاثي.

RES : ويمثل تصنيف منطقة السكن (حضري/ ريفي RES1) ومقياسه إسمي ثنائي.

BAD : ويمثل سكن البادية (خارج البادية / البادية BAD1) ومقياسه إسمي ثنائي.

WEL : ويمثل الوضع المادي (ليست فقيرة / فقيرة WEL1) ومقياسه إسمي ثنائي.

CONT : ويمثل إستخدام طرق موانع الحمل (كلا/ طرق تقليدية CONT1 / طرق حديثة CONT2) ومقياسه إسمي ثلاثي.

SMOK : ويمثل التدخين (كلا / نعم SMOK1) ومقياسه إسمي ثنائي.

WORK : ويمثل صفة العمل (لا تعمل / تعمل WORK1) ومقياسه إسمي ثنائي.

وتم تطبيق نموذج الإنحدار اللوجستي المتعدد التالي:

$$\text{Logit } Y = \log\left(\frac{P}{1-P}\right)$$

$$= \beta_0 + \beta_1 AGE + \beta_2 ED1 + \beta_3 ED2 + \beta_4 PAR1 + \beta_5 PAR2 + \beta_6 REG1 + \beta_7 REG2 + \beta_8 RES1 + \beta_9 BAD1 + \beta_{10} WEL1 + \beta_{11} CONT1 + \beta_{12} CONT2 + \beta_{13} SMOK1 + \beta_{14} WORK1$$

$$P = \text{Prob}(y=1|X) = \text{Prob}(\text{BMI} \geq 30\ \text{Kg/m}^2)$$

و X هنا تمثل مجموعة المتغيرات التوضيحية (المستقلة).

ملاحظة:

جدير بالتنويه أن أي متغير ظهر في النموذج أعلاه (عدا العمر AGE لكونه كمي) يأخذ القيمة (1) عند تحقق وجوده والقيمة (0) عدا ذلك في حال كونه ثنائي. أما في حالة كونه ثلاثي (ولنأخذ متغير مستوى التعليم ED على سبيل المثال) فإن القيم التي تحتسب لمجاميعه تكون بالشكل التالي:

$ED1$ يأخذ القيمة (1) في حالة المستوى ثانوي والقيمة (0) عدا ذلك

$ED2$ يأخذ القيمة (1) في حالة المستوى جامعي والقيمة (0) عدا ذلك

وفي حالة كون كليهما بالقيمة (0) فيعني أن مستوى التعليم بدون أو إبتدائي

وهكذا بالنسبة لبقية المتغيرات الثلاثية.

وجدير بالذكر أن الدراسة موضوع البحث إعتمدت بيانات من مسوحات ديموغرافية وصحية لثلاثة سنوات 2002 و 2009 و 2012 للنساء بعمر (49 – 15) سنة ولحالتين منهن (جميع النساء المشمولات بالمسح / النساء المتزوجات فقط). كما تم تطبيق الدراسة لكل مسح على حدة بالإضافة إلى دمج المسوحات الثلاثة.

ولغرض التوضيح فيما نتناوله من تحليل لنتائج النموذج اللوجستي، سنكتفي بعرض نتائج حالة مسح عام 2009 وللنساء المتزوجات بعمر (49 – 15) سنة نظراً لشموليته لجوانب عديدة من التحليل لنتائج هذا النموذج. والنتائج كانت حسبما مبين في الجدول التالي:

المتغير	مسح عام 2009 للمتزوجات بعمر(49 – 15)	
	β	OR
AGE	0.077 ***	1.08
ED1	– 0.041	0.96
ED2	– 0.562 ***	0.57
PAR1	0.166	1.18
PAR2	0.56 **	1.75
REG1	0.336 **	1.40
REG2	0.531 ***	1.70
RES1	0.068	1.07
BAD1	0.077	1.08
WEL1	0.00	1.00
CONT1	– 0.329 *	0.72
CONT2	– 0.117	0.98
SMOK1	– 0.528 **	0.59
WORK1	0.02	1.02

*P-value < 0.05 ** P-value < 0.01 *** P-value < 0.001

تحليل النتائج

سنتناول أولاً المتغيرات النوعية:

1) بالنسبة للمتغيرات غير المعنوية

في حالة المتغير الثلاثيED "مستوى التعليم" (بدون أو إبتدائي/ ثانوية ED1 /جامعية ED2) نجد أن ED1 والذي يمثل المستوى الثانوي غير معنوي، وهذه النتيجة تشير إلى أنه عند مقارنة

امرأتين من حيث مقياس البدانة إحداهما بمستوى تعليم إبتدائي أو دون ذلك والأخرى بمستوى تعليم ثانوي، مع تثبيت بقية المتغيرات لهما فإننا نتوقع مقياس بدانة لهما بنمط واحد أي كلتاهما فوق 30 أو كلتاهما دون ذلك. بمعنى آخر، فإن المرأة بمستوى تعليم ثانوي لا تختلف عن أخرى بمستوى تعليم إبتدائي تجاه مقياس البدانة عندما تكونا متساويتان بالنسبة للمتغيرات الأخرى.

وبنفس الوقت، لو نظرنا لحالة المتغير الثنائي RES "تصنيف منطقة السكن" (حضري/ريفي RES1) نجد أن RES1 والذي يمثل الريف غير معنوي، وهذه النتيجة تشير أيضاً إلى أنه عند مقارنة امرأتين من حيث مقياس البدانة إحداهما من سكنة منطقة حضرية والأخرى من سكنة منطقة ريفية، مع تثبيت بقية المتغيرات لهما فإننا نتوقع مقياس بدانة لهما بنمط واحد أي كلتاهما فوق 30 أو كلتاهما دون ذلك. بمعنى آخر، فإن المرأة من سكنة المنطقة الحضرية لا تختلف عن أخرى من سكنة المنطقة الريفية تجاه مقياس البدانة عندما تكونا متساويتان بالنسبة للمتغيرات الأخرى. أي أنهما يتعرضان للبدانة بإحتمال متقارب جداً.

وهكذا لبقية المتغيرات غير المعنوية.

2) بالنسبة للمتغيرات المعنوية بإشارة سالبة إلى تقدير β

ولنأخذ متغير المستوى التربوي ED على سبيل المثال ومقياسه رتبوي ثلاثي (أساسي أو دون ذلك/ ثانوي ED1 / عالي ED2). وأن ED2 الذي يمثل مستوى التعليم العالي وذو معنوية عالية جداً حيث أن (P-value < 0.001). وهذا يشير إلى أن المرأة التي لديها مستوى تعليم عالي يتوقع أن تكون أقل عرضة للبدانة مقارنة مع التي لديها مستوى تعليم أساسي أو دون ذلك عند تماثل جميع المتغيرات الأخرى للمرأتين. وبشكلٍ أدق، فإن قيمة نسبة الأرجحية (OR = 0.57) تعني أن المرأة التي لديها مستوى تعليم عالي ينخفض لديها إحتمال البدانة إلى 0.57 من نفس الإحتمال مع تلك التي لديها مستوى تعليم أساسي أو دون ذلك. أي أن هذا الإحتمال سينخفض إلى النصف تقريباً مع زيادة مستوى التعليم على نحو الشكل المبين هنا. وفي ضوء التفسير الآخر لهذه القيمة فإن مقابل كل إمرأة بدينة ضمن مجموعة المستوى التعليمي العالي سنتوقع أن نجد إمرأتان تقريباً بهذا الوضع ضمن مجموعة المستوى التعليمي الأساسي أو أقل.

3) بالنسبة للمتغيرات المعنوية بإشارة موجبة إلى تقدير β

وهنا علينا أن نتوقع إتجاهاً عكسياً لما توصلنا اليه في الفقرة (2) أعلاه.

لنأخذ المتغير PAR ويمثل حجم الولادات ومقياسه فئوي ثلاثي (حد أعلى 1 /(2-3) PAR1 / (4 ≤) PAR2). *وأن PAR2 والذي يمثل عدد الأطفال (4 ≤) وذو معنوية عالية حيث أن* (P-value < 0.01). وهذا يشير إلى أن المرأة التي لديها أربعة أطفال فأكثر يتوقع أن تكون أكثر عرضة للبدانة مقابل التي لديها طفل واحد في الأكثر عند تماثل جميع المتغيرات

الأخرى للمرأتين. وبشكلٍ أدق، فإن قيمة نسبة الأرجحية (OR = 1.75) تعني أن المرأة التي لديها طفل واحد أو بدون وقد تكون بدينة بإحتمال معين، سيزداد إحتمال أن تصبح بدينة بمقدار 75% عندما يكون لديها أربعة أطفال. أي أن هذا الإحتمال سيتضاعف تقريباً مع زيادة عدد الأطفال. وتفسيراً آخر لهذه القيمة يفيد بأنه مقابل كل إمرأة بدينة ضمن مجموعة ذوي طفل واحد أو بدون سنتوقع إمرأتان تقريباً تتسمان بالبدانة ضمن مجموعة ذوي أربعة أطفال فأكثر.

لكننا، من جانب آخر، نلاحظ عدم معنوية PAR1 والذي يمثل حالة الأمومة لطفلين أو ثلاثة مع نسبة أرجحية (OR = 1.18) والتي تعني أن المرأة التي لديها طفل واحد أو بدون وقد تكون بدينة بإحتمال معين، سيزداد إحتمال أن تصبح بدينة بمقدار 18% فقط عندما يكون لديها طفلان أو ثلاثة. وهذا يعتبر صغيراً لا تأثير له بتفسير المعنوية.

وفي مثل حالة هذا المتغير حيث أحد المستويات التعليمية (العالي ED2) له تأثير معنوي واضح والآخر (الثانوي ED1) بتأثير لا معنوي، فإننا لا نحتاج إلى المقارنة بينهما من خلال مقياس الخطورة النسبية (RR) والذي يستساغ أحياناً استخدامه للمقارنة ما بين خطورة مستويين وبشكلٍ معنوي تجاه المتغير المعتمد نسبة إلى مستوى المقارنة. ولتوضيح هذا الجانب، دعنا نأخذ المتغير REG والذي يمثل المنطقة الجغرافية (المركز / شمال REG1 / جنوب REG2) ومقياسه إسمي ثلاثي. والمركز هنا، بطبيعة الحال، يمثل مستوى المقارنة. ومن الجدول أعلاه، نجد أن نسبة الأرجحية OR لكل من الشمال (REG1) والجنوب (REG2) هي 1.40 بمستوى معنوية ($P\text{-value} < 0.01$) و 1.70 بمستوى معنوية ($P\text{-value} < 0.001$) على التوالي. وبذلك فإن الجنوب يعتبر أكثر خطورة من الشمال تجاه حدوث البدانة مقارنة بالمركز. ولكي نقارن بين خطورة هاتين المنطقتين، فإننا نحسب الخطورة النسبية RR لهما وحسب ما يلي:

$$RR = 1.70/1.40 = 1.22$$

ومعنى ذلك أن منطقة الجنوب تعتبر، تقديرياً، 1.22 مرة أكثر خطورة لحدوث البدانة من الشمال مقارنة بالمركز.

تحليل التباين متعدد المتغيرات

Multivariate Analysis of Variance (MANOVA)

إن تحليل التباين متعدد المتغيرات MANOVA عبارة عن متعدد متغيرات تعميمي لحالة تحليل التباين الأحادي ANOVA والذي هو عبارة عن اسلوب يستخدم لمقارنة أوساط عدد من المجتمعات حول متغير مقاس واحد. وعندما تكون هنالك عدد من المتغيرات المقاسة على كل وحدة تجريبية، فإنه بإمكاننا تطبيق إسلوب ANOVA بشكلٍ منفرد على كل واحد من هذه المتغيرات. على سبيل المثال، لو كان هنالك 15 متغير فإن بإمكان الباحث تنفيذ 15 تحليلاً منفصلاً بواقع تحليل ANOVA واحد منفرد لكل متغير. وعلى أية حال، هذا ليس معقولاً ولكن هذا ما نلاحظه في الغالب.

بالنسبة للإحصائيين، فإنه لديهم اعتراضين رئيسيين تجاه التحليل الإنفرادي هذا وهما:

1) إن هذه المجتمعات قد تكون مختلفة بالنسبة لبعض المتغيرات ولكنها ليست كذلك بالنسبة لمتغيرات أخرى. والباحث هنا يواجه حيرة حول قرارٍ منطقي في أي من هذه المجتمعات يمكن إعتبارها مختلفة وأياً منها متجانس. إن تحليل التباين متعدد المتغيرات يمكن أن يساعد الباحث حينها في إجراء المقارنة بين هذه المجتمعات آخذاً في الإعتبار جميع المتغيرات المشمولة بالتحليل في نفس الوقت وبعملية واحدة وبالتالي اتخاذ القرار المنطقي المناسب بشأن حقيقة إختلاف هذه لمجتمعات من عدمها.
2) ليست هنالك حماية (صيغة) كافية تجاه عمل الأخطاء من النوع الأول Type I Error عند عمل التحليل واحداً بعد آخر وبشكلٍ منفرد (تذكر من مبادئ الإحصاء كون الخطأ من النوع الأول يحدث عند رفض فرضية العدم H_0 وهي صحيحة). وهذا ناتج عن حقيقة أنه كلما زاد عدد المتغيرات لدى الباحث للتحليل، كلما زادت امكانية ظهور واحد في الأقل من هذه المتغيرات ليعطي زيادة في مستوى الثقة الإحصائية Statistical Significance. بمعنى آخر، كلما زاد عدد المتغيرات الخاضعة للتحليل كلما زاد إحتمال ظهور أحد التحليلات بفرقٍ معنوي $\{Pr(\ p\text{-value} < 0.05) \rightarrow 1\}$

بطبيعة الحال، على الباحث أن يكون حذراً لتفادي حالة زيادة احتمال الخطأ من النوع الأول. وعليه أن يكون أكثر ثقة عندما يدعي (يستنتج) بأن مجتمعين أو أكثر لديها أوساط حسابية مختلفة بالنسبة لأحد المتغيرات وأن لا يكون أحداً غيره قد يستنتج عكس ذلك من خلال تطبيق نفس التحليل وعلى نفس البيانات.

إنه من الضروري تطبيق تحليل التباين متعدد المتغيرات متى ما كانت هناك مقارنة ما بين مجتمعين أو أكثر على أساس عدد كبير من المتغيرات. فإذا ما أظهر تحليل التباين متعدد المتغيرات فروقات معنوية فإن الباحث يكون واثقاً من أن هذه الفروقات حقيقية.

أما في حالة كون تحليل التباين متعدد المتغيرات لم يظهر أية فروقات معنوية، فإن على الباحث أن يكون شديد الحذر في إعطاء أي استنتاج عند عمل التحليل المنفرد لكل متغير لأن هذه التحليلات قد لا تحدد أي شئ أكثر من " إيجابية كاذبة " false positive.

<u>T مقابل T^2</u>

لكي نرى طبيعة تحليل MANOVA، دعنا نبدأ من نقطة الصفر في هذا الإتجاه والذي موضوعه الرئيسي إيجاد إحصاءة إختبار للفرضيات حول الأوساط الحسابية للمجتمعات Population Means وبإختلاف صيغها من ثنائية إلى أكثر من ذلك.

<u>إختبار t</u>

ويستخدم إختبار Studentized t لإختبار الفرضية :

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2 \quad or \quad \mu_1 > \mu_2 \quad or \quad \mu_1 < \mu_2$$

ولغرض تسلسل الأفكار وتسهيل عملية الربط، فإننا نستخدم $H_1 : \mu_1 \neq \mu_2$ حيث أن μ_i هي الوسط الحسابي الحقيقي للمجتمع (i). ولإجراء هذا الإختبار نستخدم إحصاءة الإختبار:

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathrm{t}_{n_1+n_2-2}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

وتسمى S_p^2 بالتباين المدمج Pooled Variance وأن $\bar{X}_1$ هي الوسط الحسابي لعينة بحجم n_1 من المجتمع الأول و $\bar{X}_2$ هي الوسط الحسابي لعينة بحجم n_2 من المجتمع الثاني وليس من الضرورة أن تكون $n_1 = n_2$. كما أن S_1^2 و S_2^2 هما التباين من العينتين وبمثابة التقدير للتباين σ_1^2 و σ_2^2 للمجتمعين.

وعلينا أن نتذكر بأن كلاً من μ_1 و μ_2 هما بقيمة واحدة (أي 1x1).

والآن إذا ما كان أياً من المجتمعين يتضمن مجموعتين أو أكثر، فهذا يعني أن وسطي المجتمعين سيكونا $\underline{\mu_1}$ و $\underline{\mu_2}$ وكلٍ منهما عبارة عن متجه وكأن يكون بحجم (px1).

أي أن:

$$\underline{\mu}_1 = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{bmatrix} \quad , \quad \underline{\mu}_2 = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{bmatrix}$$

وبالتالي فإننا بصدد إختبار الفرضية:

$$H_0 : \underline{\mu}_1 = \underline{\mu}_2$$
$$H_1 : \underline{\mu}_1 \neq \underline{\mu}_2$$

وهنا يصبح لدينا موضوع تعدد المتغيرات. وبالتالي، فإن إختبار t لم يعد ممكناً إستخدامه هنا وإنما نستخدم إختبار هوتلنك Hotelling T^2 والذي هو ما يقابل إختبار t عندما تكون الأوساط بصيغة متجهات (px1).

فبينما تكون الحالة الأولى بمتغير منفرد لكلا المجتمعين مفترضين

$$X_1 \sim N(\mu_1, \sigma_1^2)$$
$$X_2 \sim N(\mu_2, \sigma_2^2) \quad , \quad \sigma_1^2 = \sigma_2^2 = \sigma^2$$

نجد في حالة متعدد المتغيرات أن:

$$\underline{X}_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1p} \end{bmatrix} \quad , \quad \underline{X}_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2p} \end{bmatrix}$$

ولذلك فإننا نفترض:

$$\underline{X}_1 \sim N_p(\underline{\mu}_1, \Sigma_1)$$
$$\underline{X}_2 \sim N_p(\underline{\mu}_2, \Sigma_2) \quad , \quad \Sigma_1 = \Sigma_2$$

$$\Sigma_1 = \left(\sigma_{ij}^{(1)}\right)_{p.xp} \quad , \quad \Sigma_2 = \left(\sigma_{ij}^{(2)}\right)_{p.xp}$$

$$\left(\sigma_{ij}^{(1)}\right)_{p.xp} = \left(\sigma_{ij}^{(2)}\right) = \Sigma_{pxp}$$

وبعد أن نحسب مصفوفة التباين المدمج $\hat{\Sigma}$ وحسب الصيغة:

$$\hat{\Sigma} = \frac{(n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2}{n_1 + n_2 - 2}$$

فإنه بالإمكان إجراء إختبار للفرضية أعلاه بإستخدام إحصاءة الإختبار T^2 بالصيغة التالية:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \left(\overline{X}_1 - \overline{X}_2\right)' \hat{\Sigma}^{-1} \left(\overline{X}_1 - \overline{X}_2\right)$$

حيث أن $\overline{X}_1$ و $\overline{X}_2$ هما تقديران لوسطي المجتمعين $\underline{\mu}_1$ و $\underline{\mu}_2$ على التوالي وأن:

$$\overline{X}_1 = \begin{bmatrix} \overline{X}_{11} \\ \overline{X}_{12} \\ \vdots \\ \overline{X}_{1p} \end{bmatrix} \quad , \quad \overline{X}_2 = \begin{bmatrix} \overline{X}_{21} \\ \overline{X}_{22} \\ \vdots \\ \overline{X}_{2p} \end{bmatrix}$$

ولغرض التبسيط في إستكمال قرار الرفض أو القبول للفرضية : $H_0 : \underline{\mu_1} = \underline{\mu_2}$ ، فإنه بالإمكان إستخدام توزيع F لهذا الغرض وعلى النحو التالي:

رفض H_0 إذا كان:

$$\frac{(n_1 + n_2 - p - 1)}{p(n_1 + n_2 - 2)} T^2 > f_{p\,,\,(n1+n2-p-1)}$$

MANOVA مقابل ANOVA

من الممكن إستخدام ANOVA بدلاً من اختبار t في حالة وجود مجتمعين أو أكثر قيد فرضية الإختبار وأن حجم μ_i هو (1x1) لجميع المجتمعات. أي أننا نختبر الفرضية:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \quad , \; k \geq 2$$
$$H_1 : at\ least \;\; \mu_i \neq \mu_j \quad , \;\; i \neq j$$

مفترضين أن: $\sigma_i^2 = \sigma^2 \;\; , \;\; i = 1, 2, ..., k$

وعملية الإختبار هذه تتم من خلال جدول تحليل التباين الآتي:

S.V.	df	SS	MS	F
Between Within(Error)	k-1 N -k	SSB SSE	MSB=SSB/(k-1) MSE=SSE/(N-k)	MSB/MSE
Total	N -1	TSS		

حيث أن $N = \sum_{i=1}^{k} n_i$ وأن n_i هو حجم العينة من المجتمع (i). ومكونات الجدول أعلاه تحتسب بالشكل التالي:

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{X_{..}^2}{N}$$

$$X_{..} = \sum_i \sum_j X_{ij}$$

$$SSB = \frac{\sum_i X_{i.}^2}{n_i} - \frac{X_{..}^2}{N}$$

$$SSE = TSS - SSB$$

والبيانات التي نتعامل معها ستكون بالشكل التالي:

Populations (groups)			
1	2	……	k
X_{11} X_{12} . . X_{1n1}	X_{21} X_{22} . . X_{2n2}	…… …… ……	X_{k1} X_{k2} . . X_{knk}
$X_{1.}$	$X_{2.}$	……	$X_{k.}$

ملاحظة:

في حالة المجموعتين (k=2) فإنه يمكن إستخدام إحصاءة الإختبار F أو T وفي هذه الحالة فإن القيم الحسابية $T^2 = F$ وهذه تقودنا إلى مقارنة القيم الجدولية

$$t_{\alpha/2\ ,(n1+n2-2)}^2 = f_{\alpha\ ,(1\ ,\ n1+n2-2)}$$

وبإختصار، فإذا كانت ANOVA تتعامل مع اختبار الفروقات ما بين أوساط حسابية للمجتمعات (إثنان أو أكثر) فإن MANOVA تتعامل مع اختبار الفروقات ما بين متجهات الأوساط الحسابية (إثنان أو أكثر).

بمعنى آخر، وبلغة المصفوفات، فإن ANOVA تتعامل مع متجه أوساط حسابية بأبعاد (1x1) لأي مجموعة (مجتمع)، بينما MANOVA تتعامل مع متجه أوساط حسابية بأبعاد (px1) لأي مجموعة (مجتمع) وأن p يمثل عدد المتغيرات المعتمدة الناتجة من خلال التجربة.

ملاحظة:

يجدر بنا أن نعلم بأن أياً من هاتين الطريقتين لا تعطينا إستنتاجاً مباشراً عن أي من هذه الأوساط سيكون مختلفاً عن الآخر في حالة تثبيت وجود فروقات بشكلٍ معنوي ما بين هذه الأوساط. وللتغلب على هذه الإشكالية في الطريقتين، يمكننا إستخدام أساليب أخرى لهذا الغرض.

ومن أجل التوضيح، فإننا نستخدم ANOVA لإختبار الفرضية:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

حيث كل μ_i عبارة عن قيمة مفردة .(1x1)

بينما نستخدم MANOVA لإختبار الفرضية:

$$H_0 : \underline{\mu}_1 = \underline{\mu}_2 = ... = \underline{\mu}_k$$

حيث كل $\underline{\mu}_i$ عبارة عن قيمة مفردة (px1). أي أن:

$$\underline{\mu}_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{ip} \end{bmatrix}$$

وأن: k تمثل عدد المجاميع (المجتمعات).

P هي حجم المتجه والذي يعكس عدد المتغيرات المعتمدة التي يتم قياسها خلال التجربة.

كيفية إختبار الفرضية H_0

كما نتذكر الإختبار لهذه الفرضية في حالة ANOVA حيث أن TSS للمتغير المعتمد يتم تقسيمه إلى SSB و SSE. ومع التعدد في المتغير المعتمد وإعتماد إسلوب MANOVA ، فإنه لا زال بالإمكان احتسابSSB و SSE والتي ستكون عبارة عن مصفوفة (pxp). بالإضافة إلى ذلك، فإنه بالإمكان إحتساب مجموع الضرب المتبادل (المتقاطع) Sum of Cross Products وتجزئته إلى SSB و SSE المتجمع Pooled. وهنا في حالة MANOVA، لعله من المناسب توصيف الآتي من المصفوفات (مفترضين متغيرين معتمدين X_1 و X_2 لغرض التوضيح الأسهل لهذه العملية):

$$W = \text{Pooled within groups (SSCP)} = \begin{bmatrix} SSW_1 & SCP_w \\ SCP_w & SSW_2 \end{bmatrix}$$

والرموز هذه تعني:

SSCP = (Sum of Squares for Cross Product) مجموع مربعات الضرب المتبادل

مجموع المربعات المدمج ضمن:

SSW_1 = Pooled SS within groups for X_1

SSW_2 = Pooled SS within groups for X_2

SCP_w = Pooled within group Sum of Product for X_1 & X_2

$$B = \begin{bmatrix} SSb_1 & SCP_b \\ SCP_b & SSb_2 \end{bmatrix}$$

SSb_i = Between – groups SS for X_i

SCP_b = Between – groups Sum of Cross Products of X_1 & X_2

$$T = \begin{bmatrix} SS_1 & SCP_{12} \\ SCP_{12} & SS_2 \end{bmatrix}$$

SS_i = The Total SS for Xi

SCP_{12} = The Total Sum of Cross Products of X_1 & X_2

ومن المعلوم أن T = B + W (Total = Between + Within).

ولأجل توضيح كيفية إحتساب هذه العناصر للمصفوفات T , B , W دعنا نفترض مجموعتين G1 , G2 ومتغيرين معتمدين لكلٍ منهما X_1 و X_2 بالبيانات التالية :

G1		G2	
X_1	X_2	X_1	X_2
8	3	4	2
7	4	3	1
5	5	3	2
3	4	2	2
3	2	2	5
$\sum X = 26$	18	14	12
$\sum X^2 = 156$	70	42	38
$\sum X_1X_2 =$ 95		31	

$$\sum X_{t1} = 26 + 14 = 40$$
$$\sum X_{t2} = 18 + 12 = 30$$
$$\sum X_{t1}^2 = 156 + 42 = 198$$
$$\sum X_{t2}^2 = 70 + 38 = 108$$
$$CP_t = 95 + 31 = 126$$

ومن البيانات هذه يمكننا حساب العناصر التالية:

$$SSW_1 = \left[156 - \frac{(26)^2}{5}\right] + \left[42 - \frac{(14)^2}{5}\right] = 23.6$$
$$SSW_2 = \left[70 - \frac{(18)^2}{5}\right] + \left[38 - \frac{(12)^2}{5}\right] = 14.4$$
$$SCP_w = \left[95 - \frac{(26)(18)}{5}\right] + \left[31 - \frac{(14)(12)}{5}\right] = -1.2$$

$$W = \begin{bmatrix} 23.6 & -1.2 \\ -1.2 & 14.4 \end{bmatrix}$$

كذلك فإن:

$$SSb_1 = \left[\frac{(26)^2}{5} + \frac{(14)^2}{5}\right] - \frac{(40)^2}{10} = 14.4$$
$$SSb_2 = \left[\frac{(18)^2}{5} + \frac{(12)^2}{5}\right] - \frac{(30)^2}{10} = 3.6$$
$$SCP_b = \left[\frac{(26)(18)}{5} + \frac{(14)(12)}{5}\right] - \frac{(40)(30)}{10} = 7.2$$

$$B = \begin{bmatrix} 14.4 & 7.2 \\ 7.2 & 3.6 \end{bmatrix}$$

وبالتالي وعن طريق الجمع نجد أن:

$$T = B + W = \begin{bmatrix} 14.4 & 7.2 \\ 7.2 & 3.6 \end{bmatrix} + \begin{bmatrix} 23.6 & -1.2 \\ -1.2 & 14.4 \end{bmatrix} = \begin{bmatrix} 38 & 6 \\ 6 & 18 \end{bmatrix}$$

ملاحظة:

جدير بالذكر أن عناصر المصفوفة T يمكن احتسابها بشكلٍ مستقل وعلى النحو التالي:

$$SST_1 = 198 - \frac{(40)^2}{10} = 38$$

$$SST_2 = 108 - \frac{(30)^2}{10} = 18$$

$$SCP_t = (95 + 31) - \frac{(40)(30)}{10} = 6$$

ولأننا نعمل على مجموعتين، فإن الفرضية التي نحن بصدد عمل الإختبار لها هي:

$$H_0 : \underline{\mu}_1 = \underline{\mu}_2$$

وأن إحصاءة الإختبار العامة في حالة المجموعتين هي:

$$F = \frac{(1 - \Lambda)/t}{\Lambda/(N - t - 1)} \sim f_{t\,,\,(N-t-1)}$$

حيث $t = 2$ وهي عدد المتغيرات المعتمدة X_1 و X_2

$N = 10$ المجموع الكلي لعدد المشاهدات لكل مجموعة

والرمز Λ يشير إلى ما يعرف (ولكس لامبدا Wilk's Lambda)

$$\Lambda = \frac{|W|}{|T|} = \frac{338.4}{648} = 0.52222$$

وحيث أن مثالنا هذا فيه $N=10$ و $t=2$ فإن:

$$F = \frac{(1 - 0.52222)/2}{0.52222/(10 - 2 - 1)} = 3.202$$

وهذه القيمة المحسوبة لإحصاءة الإختبار هي أقل من القيمة الجدولية المقابلة لها وهي:

$$f_{t\,,\,(N-t-1)} = f_{2\,,\,7}\,(0.05) = 4.74$$

فإننا لا نرفض الفرضية H_0 ونستنتج أن لا فروقات معنوية بين المجموعتين بالنسبة إلى أوساط مجموعة المتغيرات المعتمدة X_1 و X_2.

<u>تنويه:</u>

إن ما أوردناه ضمن هذا المثال لا يمثل الحالة العامة لإحتساب الإحصاءة F والتي سيتم تناولها تالياً.

الحالة العامة لإحتساب الإحصاءة F

لو إفترضنا الأبعاد التالية في التجربة وهي:

K = عدد المجاميع

P = عدد المتغيرات المعتمدة في كل مجموعة ومقياسها كمي (نسبي أو فئوي)

N = العدد الكلي للمشاهدات

وفي ضوء ذلك، ستكون قيمة الإحصاءة F حسب الصيغة التالية

$$F = \frac{\left(1 - \Lambda^{1/b}\right) / df_1}{\Lambda^{1/b} / df_2}$$

حيث أن:

$$df_1 = P(K-1)$$
$$df_2 = ab - c$$
$$a = N - K - \frac{P - K + 2}{2}$$
$$b = \sqrt{\frac{P^2(K-1)^2 - 4}{P^2 + (K-1)^2 - 5}}$$
$$c = \frac{P(K-1) - 2}{2}$$

مع ملاحظة أن قيمة $b = 1$ في حالة عدم تحقق كون ($P^2 + (K-1)^2 - 5 > 0.0$).

وإذا ما طبقنا هذه الصيغة العامة على المثال أعلاه حيث (K=2 , P=2 , N=10)

نجد أن (df_1= 2 , df_2 = 7 , c =0 , b =1 , a = 7)

وبالتالي فإن الحسابات التي اعتمدت تكون متوافقة كلياً مع الحالة العامة.

وسوف نرى أبعاد تطبيق الحالة العامة من خلال المثال التالي لبيانات حقيقية:

مثال:

في تجربة زراعية لمعرفة ما إذا توجد فروقات معنوية ما بين 4 نوعيات من التربة (سطحية، رملية، ملحية، طينية) بالنسبة لزراعة نوع جديد من بذور الذرة في ضوء 3 من

المتغيرات المعتمدة (الناتج، كمية مياه الري، كمية المبيد الحشري) واستخدام 8 نباتات لكل نوع من التربة[5].

K = عدد المجاميع = 4

P = عدد المتغيرات المعتمدة = 3

N = العدد الكلي للمشاهدات = 32

وبالتالي سيكون لدينا، وفقاً للصيغ أعلاه، القيم التالية:

$$a = N - K - \frac{P-K+2}{2} = 32 - 4 - \frac{3-4+2}{2} = 27.5$$

$$b = \sqrt{\frac{P^2(K-1)^2 - 4}{P^2 + (K-1)^2 - 5}} = \sqrt{\frac{3^2(4-1)^2 - 4}{3^2 + (4-1)^2 - 5}} = 2.434$$

$$c = \frac{P(K-1)-2}{2} = \frac{3(4-1)-2}{2} = 3.5$$

$$df_1 = P(K-1) = 3(4-1) = 9$$

$$df_2 = ab - c = (27.5)(2.434) - 3.5 = 63.43$$

ومن خلال حسابات مصفوفات التباين والتباين المشترك الثلاثة W و B و T وجدنا الآتي:

$$W = \begin{bmatrix} 4058 & 714 & -273 \\ 714 & 2834 & 123 \\ -273 & 123 & 113 \end{bmatrix}$$

$$B = \begin{bmatrix} 911 & 63 & 163 \\ 63 & 122 & 24 \\ 163 & 24 & 32 \end{bmatrix}$$

$$T = \begin{bmatrix} 4969 & 777 & -110 \\ 777 & 2956 & 147 \\ -110 & 147 & 145 \end{bmatrix}$$

وبالتالي نجد أن:

$$\Lambda = \frac{|W|}{|T|} = 0.489$$

$$F = \frac{1-(0.489)^{1/2.434}}{(0.489)^{1/2.434}}\left(\frac{63.43}{9}\right) = 2.405 > f_{9,\,63.43}\ (0.05) = 2.032$$

وبالتالي يتم رفض الفرضية:

$$H_0 : \underline{\mu_1} = \underline{\mu_2} = \underline{\mu_3} = \underline{\mu_4}$$

مستنتجين بوجود فروق معنوية ما بين متجهات الأوساط الحسابية الأربعة والتي يتضمن كلاً منها ثلاثة أوساط حسابية للمتغيرات المعتمدة. أي أن متجهات أوساط (الناتج، مياه الري، مبيد الحشرات) تختلف بشكلٍ معنوي ما بين نوعيات التربة (سطحية، رملية، ملحية، طينية).

التحليل المميز
Discriminant Analysis (DA)

إن التحليل المميز يستخدم بشكلٍ رئيسي لتصنيف الأفراد أو الوحدات التجريبية إلى إثنين أو أكثر من المجتمعات المحددة بشكلٍ منفرد لا تداخل فيما بينها. ولأجل تطوير قاعدة تمييز لغرض تصنيف Classifying الوحدات التجريبية لواحدة من عددٍ من الفئات المحتملة، فعلى الباحث أن يحصل على عينة عشوائية من الوحدات التجريبية من كل فئة محتملة للتصنيف. وبعد ذلك، فإن التحليل المميز يهيئ طرقاً تمكن الباحث من بناء قواعد يمكن استخدامها لتصنيف وحدات تجريبية أخرى ضمن واحدة من الفئات التصنيفية.

ومثالاً على ذلك، لو أن شركة معينة متخصصة في منح بطاقات إئتمانية Credit Cards فإنها حتماً ترغب في أن تكون قادرة على تصنيف طلبات الحصول على البطاقة إلى مجموعتين من الأفراد (1- أفراداً جيدين من ناحية خطورة الإئتمان) و (2- أفراداً غير جيدين من ناحية خطورة الإئتمان). والشركة تمنح المجموعة الأولى بطاقات الإئتمان فيما لا تمنحها للمجموعة الثانية. ولأجل مساعدة الشركة في هذا الجانب وتحديد المجموعتين، فإنها قد تأخذ في الإعتبار عدد من الخصائص الديموغرافية التي يمكن قياسها لدى كل فرد. فالشركة قد تأخذ، على سبيل المثال، المستوى التعليمي، الراتب الشهري، المديونيات، سجل الماضي في الإئتمان كمؤشرات تنبؤية بشأن استحقاق البطاقة. والشركة بعد ذلك تحاول إستخدام هذه المعلومات لفردٍ محدد للمساعدة في إتخاذ قرار منحه البطاقة من عدمها. الشركة بحاجة لتوظيف طريقة من طرق التحليل متعدد المتغيرات لمساعدتها في تصنيف الأفراد لأيٍ من المجموعتين، وهذه الطريقة هي طريقة التحليل المميز.

بالنسبة لهذا المثال، على الشركة أن تجمع هذه البيانات من أفراد معروفين لديها بأنهم ينتمون للمجموعة الأولى وإعادة أخذ نفس البيانات من أفراد معروفين لديها أيضاً بأنهم ينتمون للمجموعة الثانية. ومن ثم بإستطاعة الشركة تصنيف طالبي البطاقة الجدد لأيٍ من المجموعتين بإستخدام القاعدة الناتجة من تطبيق التحليل المميز على الأفراد المعروفين لديها.

من المهم هنا أن يكون في بالنا احتمال ظهور خطأ في التصنيف وبإحتمال معين يتم تثبيته من خلال إعادة تصنيف الأفراد المعروفين لديها في ضوء تطبيق القاعدة الناتجة عليهم. ويمكن أن نحكم على جودة قاعدة التصنيف الناتجة بتناسب عكسي مع مقدار احتمال خطأ إعادة التصنيف لكامل المجموعة.

إن التحليل المميز هو أحد المواضيع المهمة في التحليل المتعدد المتغيرات multivariate analysis والذي يهتم في كيفية التمييز بين مجموعتين أو أكثر. إن الفكرة الاساسية من التمييز discriminate هو التفريق ما بين المجتمعات المتداخلة أو المتشابكة ولها نفس الخصائص أو

الصفات. بمعنى أخر، لنفرض انه لدينا مجتمعين أو اكثر ولدينا عينة تحتوي على مجموعة من المشاهدات من كل مجتمع. إن وظيفة التحليل المميز هي إيجاد دالة يمكن بواسطتها تصنيف أو تمييز المشاهدات الجديدة بالنسبة لمجتمعاتها الأصلية.

إن التحليل المميز يختلف عن تحليل الإنحدار في أن المتغير المعتمد في التحليل التمييزي هو متغيرذو مقياس إسمي nominal variable وهو من المتغيرات النوعية (يأخذ قيمتين 0 , (1) في حالة التمييز ما بين مجموعتين

$Y = 1$ إذا كانت المشاهدة تعود للمجتمع الأول.
$= 0$ إذا كانت المشاهدة تعود للمجتمع الثان.

بينما المتغير المعتمد في تحليل الإنحدار هو على الأكثر متغير مستمر (وهو من المتغيرات الكمية) ويتشابه التحليلين بأن كلاهما يهدف لإيجاد علاقة بين المتغير المعتمد والمتغيرات المستقلة.

أنواع الدوال التمييزية

1. الدالة المميزة الخطية
2. االدالة المميزة التربيعية
3. الدالة المميزة اللوجستية

1. الدالة المميزة الخطية Linear Discriminant Function

تستخدم هذه الدالة عندما تكون المجتمعات المدروسة ذات توزيع طبيعي متعدد المتغيرات بمتجهات متوسط مختلفة و مصفوفة تباين مشترك متساوية وهناك حالتان:

1- حالة مجموعتين (مجتمعين)
2- حالة عدة مجاميع (مجتمعات)

حالة المجموعتين

نفرض لدينا عينة مسحوبة من مجتمعين يتوزعان توزيعاً طبيعياً بمتوسطين $\underline{\mu_2}$, $\underline{\mu_1}$ ومصفوفة تباين مشترك Σ لكل مجتمع. أي أن:

$$\underline{X_1} \sim N_p(\underline{\mu_1}\ ,\ \Sigma\ )$$
$$\underline{X_2} \sim N_p(\underline{\mu_2}\ ,\ \Sigma\ )$$

يمكن صياغة دالة بالإعتماد على مقاييس من هذه القيم وأن هذه الدالة تمكننا من إختيار أي مشاهدة و تحديد المجتمع الذي تعود إليه. إن المتغير العشوائي X له عادةً دالة كثافة إحتمالية إما f_1

$f_1(x, \theta_1)$ أو $f_2(x, \theta_2)$ و θ_i هنا ترمز لأي معلمة يختلف التوزيع بموجبها. وبذلك، فإن هذا يعني أن:

$$f_i(x, \mu_i) = \frac{1}{(2\pi)^{\frac{p}{2}} (\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu_i})' \Sigma^{-1} (\underline{x}-\underline{\mu_i})} \quad , \quad i = 1, 2$$

وبذلك فإن نسبة الإمكان الأعظم هنا تكون:

$$\frac{f_1(x, \mu_1)}{f_2(x, \mu_2)} = \frac{\frac{1}{(2\pi)^{\frac{p}{2}} (\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu_1})' \Sigma^{-1} (\underline{x}-\underline{\mu_1})}}{\frac{1}{(2\pi)^{\frac{p}{2}} (\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu_2})' \Sigma^{-1} (\underline{x}-\underline{\mu_2})}} \geq \lambda$$

أو أن:

$$\exp\left[-\frac{1}{2}\left\{(\underline{x}-\underline{\mu_1})' \Sigma^{-1} (\underline{x}-\underline{\mu_1}) - (\underline{x}-\underline{\mu_2})' \Sigma^{-1} (\underline{x}-\underline{\mu_2})\right\}\right] \geq \lambda$$

وعند تفكيك هذه المعادلة وإضافة وطرح المقدار $\underline{\mu_2}' \Sigma^{-1} \underline{\mu_1}$ يصبح لدينا:

$$\exp\left[\underline{x}' \Sigma^{-1} (\underline{\mu_1}-\underline{\mu_2}) - \frac{1}{2}\left\{(\underline{\mu_1}+\underline{\mu_2})' \Sigma^{-1} (\underline{\mu_1}-\underline{\mu_2})\right\}\right] \geq \lambda$$

وبأخذ اللوغاريتم الطبيعي للطرفين:

$$\left[\underline{x}' \Sigma^{-1} (\underline{\mu_1}-\underline{\mu_2}) - \frac{1}{2}\left\{(\underline{\mu_1}+\underline{\mu_2})' \Sigma^{-1} (\underline{\mu_1}-\underline{\mu_2})\right\}\right] \geq \log \lambda$$

وعندما تكون $\lambda = 1$ فإن:

$$\left[\underline{x}' \Sigma^{-1} (\underline{\mu_1}-\underline{\mu_2}) - \frac{1}{2}\left\{(\underline{\mu_1}+\underline{\mu_2})' \Sigma^{-1} (\underline{\mu_1}-\underline{\mu_2})\right\}\right] \geq 0$$

وبإستخدام مقدرات الإمكان الأعظم إلى كلٍ من Σ, μ_2, μ_1 وتعويضها هنا، يصبح لدينا:

$$W = \left[\underline{x}' S^{-1} (\underline{\bar{x}_1}-\underline{\bar{x}_2}) - \frac{1}{2}(\underline{\bar{x}_1}+\underline{\bar{x}_2})' S^{-1} (\underline{\bar{x}_1}-\underline{\bar{x}_2})\right] \geq 0$$

$$\underline{\bar{x}_1} = \sum_{j=1}^{n_1} \underline{x_{1j}} / n_1$$

حيث أن:

$$\underline{\bar{x}_2} = \sum_{j=1}^{n_2} \underline{x_{2j}} / n_2$$

$$S=\left[\sum_{j=1}^{n_1}\left(\underline{x_{1j}}-\underline{\bar{x}_1}\right)\left(\underline{x_{1j}}-\underline{\bar{x}_1}\right)'+\sum_{j=1}^{n_2}\left(\underline{x_{2j}}-\underline{\bar{x}_2}\right)\left(\underline{x_{2j}}-\underline{\bar{x}_2}\right)'\right]/(n_1+n_2-2)$$

وأن جزئي المعادلة w أعلاه يتكون من دالة التمييز الخطية:

$$y=\underline{x}'S^{-1}\left(\underline{\bar{x}_1}-\underline{\bar{x}_2}\right)$$

ونقطة الفصل أو التمييز z:

$$z=\frac{1}{2}D^2=\frac{1}{2}\left(\underline{\bar{x}_1}+\underline{\bar{x}_2}\right)'S^{-1}\left(\underline{\bar{x}_1}-\underline{\bar{x}_2}\right)$$

حيث أن D^2 هي إحصاءة مهلونوبيس Mahalanobis.

وأن دالة التمييز الخطية ممكن أن تكتب على شكل الدالة الخطية التالية:

$$y=\underline{x}'\underline{C} \qquad \text{or} \qquad y=\underline{C}'\underline{x}$$

حيث:

$$C=S^{-1}\left(\underline{\bar{x}_1}-\underline{\bar{x}_2}\right)$$

والآن يمكن أن نستخلص الآتي من دالة التمييز الخطية:

$$\bar{y}_1=\underline{\bar{x}_1}'S^{-1}\left(\underline{\bar{x}_1}-\underline{\bar{x}_2}\right)$$
$$\bar{y}_2=\underline{\bar{x}_2}'S^{-1}\left(\underline{\bar{x}_1}-\underline{\bar{x}_2}\right)$$

وبالتالي يمكننا التعبير عن نقطة الفصل بالآتي:

$$z=\frac{\bar{y}_1+\bar{y}_2}{2}$$

ولو إفترضنا أن $\bar{y}_1<\bar{y}_2$ فإن خطة التصنيف للمشاهدة الجديدة y هو أنها:

تعود للمجموعة الأولى في حالة كون $y\leq z$

تعود للمجموعة الثانية في حالة كون $y>z$

أو يمكن إستخدام w أعلاه بالكامل لغرض التصنيف للمشاهدة x بكونها:

تعود للمجموعة الأولى في حالة كون $w\leq 0$

تعود للمجموعة الثانية في حالة كون $w>0$

وجدير بالذكر أن المصفوفة S تستخرج بالشكل التالي (مفترضين n = 3):

$$S = \begin{bmatrix} V_{11} & V_{12} & V_{13} \\ & V_{22} & V_{23} \\ & & V_{33} \end{bmatrix}$$

حيث أن:

$$V_{ii} = \frac{S_{ii}(1) + S_{ii}(2)}{n_1 + n_2 - 2} \quad , \quad V_{ij} = \frac{S_{ij}(1) + S_{ij}(2)}{n_1 + n_2 - 2}$$

$$S_{ii} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \quad , \quad S_{ij} = \sum x_i x_j - \frac{\left(\sum x_i\right)\left(\sum x_j\right)}{n}$$

الإختبارات المستخدمة في التحليل المميز

مع أن كفاءة الدالة المميزة الخطية تقاس وفقاً لنسبة التصنيف الصحيح للمشاهدات حسب مجاميعها الأصلية، إلا أنه من الممكن التنبؤ مسبقاً بشئ عن هذه الكفاءة لأنها تزداد وفقاً لزيادة الفرق ما بين أوساط المجاميع من جهة، وتقارب قيم مصفوفات التباين – التباين المشترك لهذه المجاميع من جهة أخرى. وفي لغة الإحصاء ومفهومه، فإن ذلك يعني تطبيق إختبارات إحصائية ونستعين بنتائجها لتقييم كفاءة الدالة المميزة الخطية. وفيما يلي الإختبارات المناسبة في هذا الجانب.

1. اختبارمعنوية الفرق بين الأوساط عن طريق إختبار الفرضية:

$$H_0 : \underline{\mu_1} = \underline{\mu_2}$$
$$H_1 : \underline{\mu_1} \neq \underline{\mu_2}$$

ويتم الإختبار بإستخدام إختبار F الذي يعتمد على إحصاءة هوتلنك Hotteling T^2 والتي تكون:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2$$

حيث أن D^2 هي إحصاءة مهلونوبيس Mahalanobis والتي صيغتها

$$D^2 = \left(\underline{\bar{x}_1} - \underline{\bar{x}_2}\right)' S^{-1} \left(\underline{\bar{x}_1} - \underline{\bar{x}_2}\right)$$

وبالتالي، فإن إختبار F سيكون:

$$F = \frac{n_1 + n_2 - p - 1}{n_1 + n_2 - 2} T^2 \sim f_{p, n_1 + n_2 - p - 1}$$

والذي نريد نتيجته الرفض بمعنوية عالية لعكس زيادة نسبة التصنيف الصحيح. حيث:

n_1 : حجم العينة الأولى
n_2 : حجم العينة الثانية
P : عدد المتغيرات

2. إختبار تساوي مصفوفتي التباين- التباين المشترك للمجموعات من خلال الفرضية:

$H_0 : \Sigma_1 = \Sigma_2$
$H_1 : \Sigma_1 \neq \Sigma_2$

وأن إحصاءة الإختبار لهذه الفرضية Q وهي:

$$Q = MC^* \sim \chi^2_{(k-1)(p-1)}$$

حيث أن:

$$M = \ln \frac{|S|^{n_1+n_2}}{|S_1|^{n_1}|S_2|^{n_2}}$$

$$\text{or} \qquad M = \sum_{i=1}^{2} n_i \ln|S| - \sum_{i=1}^{2} n_i \ln|S_i|$$

$$S = \frac{A_1 + A_2}{n_1 + n_2 - 2}$$

$$A_1 = (n_1 - 1)S_1$$

$$A_2 = (n_2 - 1)S_2$$

$$C^* = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\left[\sum \frac{1}{n_i} - \frac{1}{\sum n_i}\right]$$

وأن:

k = عدد المجاميع
p = عدد المتغيرات

إن الفرضية أعلاه والتي تنص على تساوي مصفوفتي التباين - التباين المشترك بين المجموعتين هي أهم فرضية هنا لأن عدم تحققها يعني أننا لا يمكن أن نستخدم الدالة التمييزية الخطية وإنما نستخدم الدالة التمييزية التربيعية.

احتمال خطأ التصنيف: the probability of misclassification

هو إحتمال تصنيف مشاهدة معينة إلى المجموعة الأولى بينما هي تعود في الحقيقة إلى المجموعة الثانية و بالعكس. نفترض لحساب خطأ التصنيف أن حجم العينة يكون كبير لذلك فإننا نضمن كون توزيع المشاهدات يقترب من التوزيع الطبيعي (حسب نظرية الحد المركزي). حيث

أن هذا الخطأ يعتمد على أن توزيع العينة هو التوزيع الطبيعي أو يقترب من التوزيع الطبيعي.
هذا الإحتمال يكون:

P_{12} = P(classifying x to be from group(1)/ x is from group(2))
$= \phi(-D/2)$

حيث D^2 هي إحصاءة مهالونوبيس.

ويتم ايجاد هذه القيمة من جداول التوزيع الطبيعي القياسي. إن خطأ التصنيف هو عامل مهم لإثبات كفاءة الدالة المميزة. والتي تعطي أقل خطأ تصنيف هي الدالة الأكثر كفاءة و تكون الأفضل من بين دوال التمييز.

وبالإمكان أيضاً إستخدام طريقة إعادة التعويض في هذا الجانب وكما هو في أدناه.

طريقة التعويض: Resubstitution Method

تسخدم هذه الطريقة لإيجاد إحتمال خطأ التصنيف وأن إسلوب هذه الطريقة يعتمد على أنه لو كان n_j يمثل عدد المشاهدات التي تعود للمجموعة (j) وأن n_{ij} هو عدد المشاهدات في المجموعة (j) وصنفت وفق دالة التمييزعلى انها تعود للمجموعة (i)، فإن تقدير إحتمال خطأ التصنيف في هذه الحالة وبشكلٍ عام سيكون:

$$P_{ij} = \frac{n_{ij}}{n_j}$$

وفي حالة المجموعتين التي نحن فيها الآن، فإن إحتمال خطأ التصنيف الكلي للدالة المميزة سيكون:

$$\hat{p} = \frac{n_{12} + n_{21}}{n_1 + n_2}$$

أما متوسط إحتمال خطأ التصنيف سيكون:

$$\hat{p} = \frac{\hat{p}_{21} + \hat{p}_{12}}{2}$$

حالة عدة مجاميع

في حالة كون مسألة التمييز بين أكثر من مجموعتين (k من المجاميع)، يتم التصنيف عن طريق المقارنة بين كل مجموعتين وتكون لذلك عدة دوال مميزة y_{ij} وعددها C_2^k وتكتب على النحو التالي:

$$y_{ij} = \underline{x}' S^{-1}(\underline{\bar{x}}_i - \underline{\bar{x}}_j)$$

ولو كانت لدينا ثلاثة مجاميع ولكل مجموعة p من المتغيرات ($p \geq 2$) فإنه من المناسب حساب:

$$\underline{C_1} = S^{-1}\left(\underline{\bar{x}_1} - \underline{\bar{x}_2}\right)$$
$$\underline{C_2} = S^{-1}\left(\underline{\bar{x}_1} - \underline{\bar{x}_3}\right)$$
$$\underline{C_3} = S^{-1}\left(\underline{\bar{x}_2} - \underline{\bar{x}_3}\right)$$

وبذلك يكون لدينا دوال التمييز التالية:

$$Y_{12} = \underline{x}'\underline{C_1} \quad , \; Y_{13} = \underline{x}'\underline{C_2} \quad , \quad Y_{23} = \underline{x}'\underline{C_3}$$

ثم نجد ما يلي:

$$W_{12} = \left[\underline{x}'S^{-1}\left(\underline{\bar{x}_1} - \underline{\bar{x}_2}\right) - \frac{1}{2}\left(\underline{\bar{x}_1} + \underline{\bar{x}_2}\right)' S^{-1}\left(\underline{\bar{x}_1} - \underline{\bar{x}_2}\right)\right]$$
$$W_{13} = \left[\underline{x}'S^{-1}\left(\underline{\bar{x}_1} - \underline{\bar{x}_3}\right) - \frac{1}{2}\left(\underline{\bar{x}_1} + \underline{\bar{x}_3}\right)' S^{-1}\left(\underline{\bar{x}_1} - \underline{\bar{x}_3}\right)\right]$$
$$W_{23} = \left[\underline{x}'S^{-1}\left(\underline{\bar{x}_2} - \underline{\bar{x}_3}\right) - \frac{1}{2}\left(\underline{\bar{x}_2} + \underline{\bar{x}_3}\right)' S^{-1}\left(\underline{\bar{x}_2} - \underline{\bar{x}_3}\right)\right]$$

فتكون العلاقة بينها:

$$W_{23} = W_{13} - W_{12}$$

وتكون قاعدة التصنيف إذا كانت لدينا عدد المتغيرات ($p \geq 2$) لكل مجموعة حسب الآتي:

تصنف المشاهدة x لواحدة من المجاميع التالية:

- ضمن المجموعة (1) إذا كانت $W_{12} > 0$ و $W_{13} > 0$
- ضمن المجموعة (2) إذا كانت $W_{13} > 0$ و $W_{13} > W_{12}$
- ضمن المجموعة (3) إذا كانت $W_{13} > 0$ و $W_{12} > W_{13}$

أما إذا كان لدينا متغير وحيد لكل مجموعة ($p=1$) وأن أوساط المجاميع رتبت للسهولة بالشكل التالي $\bar{X}_1 < \bar{X}_2 < \bar{X}_3$ فإن قواعد التصنيف للمشاهدة x تكون كالاتي:

- ضمن المجموعة (1) إذا كانت $x < \frac{1}{2}\left(\underline{\bar{x}_1} + \underline{\bar{x}_2}\right)$
- ضمن المجموعة (2) إذا كانت $\frac{1}{2}\left(\underline{\bar{x}_1} + \underline{\bar{x}_2}\right) \leq x \leq \frac{1}{2}\left(\underline{\bar{x}_2} + \underline{\bar{x}_3}\right)$
- ضمن المجموعة (3) إذا كانت $x > \frac{1}{2}\left(\underline{\bar{x}_1} + \underline{\bar{x}_2}\right)$

ومصفوفة S لهذه الحالة تستخرج بالشكل التالي:

$$S = \begin{bmatrix} V_{11} & V_{12} & V_{13} \\ & V_{22} & V_{23} \\ & & V_{33} \end{bmatrix}$$

$$V_{ii} = \frac{S_{ii}(1) + S_{ii}(2) + S_{ii}(3)}{n_1 + n_2 + n_3 - 3} \quad , \quad S_{ii} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$V_{ij} = \frac{S_{ij}(1) + S_{ij}(2) + S_{ij}(3)}{n_1 + n_2 + n_3 - 3} \quad , \quad S_{ij} = \sum x_i x_j - \frac{\sum x_i \sum x_j}{n}$$

2. دالة التميز التربيعية Quadratic Discriminant Function

وكما ذكرنا سابقاً، يستخدم هذا النوع من الدوال في حالة عدم تساوي مصفوفة التباين - التباين المشترك بين المجموعات. و تقدر معالم هذه الدالة بطريقة MLE بافتراض أن حجم العينة كبير بحيث يصبح من الممكن أن نفترض بأن المشاهدات تقترب من التوزيع الطبيعي (النظرية المركزية).

وفي حالة المجتمعين، فإن تقدير μ_1 هو $\underline{\bar{X}_1}$ و أن تقدير μ_2 هو $\underline{\bar{X}_2}$ وتقدير Σ_1 هو S_1 وتقدير Σ_2 هو S_2. وأن مقياس التمييز هو V حيث

$$V = \frac{f_1(x_i)}{f_2(x_i)} > or < 1$$

$$G = \ln V = \ln f_1(x_i) - \ln f_2(x_i) > or < 0$$

<u>ملاحظة:</u>

في حالة وجود p من المتغيرات لكل مجموعة فإن G هذه بطبيعة الحال ستكون:

$$G = \ln f_1(x_1, x_2, ..., x_p) - \ln f_2(x_1, x_2, ..., x_p)$$

ووفقاً لما جاء في أعلاه، فإن دالة التمييز التربيعية التقديرية في حالة متغيرين ستكون:

$$\hat{G} = \frac{1}{2}\ln\frac{S_2}{S_1} - \frac{1}{2}\left(\underline{\bar{x}_1'}S_1^{-1}\underline{\bar{x}_1} - \underline{\bar{x}_2'}S_2^{-1}\underline{\bar{x}_2}\right) + \underline{x'}\left(S_1^{-1}\underline{\bar{x}_1} - S_2^{-1}\underline{\bar{x}_2}\right) - \frac{1}{2}\underline{x'}\left(S_1^{-1} - S_2^{-1}\right)\underline{x}$$

وبالتالي، فإن التصنيف للمشاهدة x سيكون

- تعود للمجموعة الأولى إذا كانت $\hat{G} > 0$
- تعود للمجموعة الثانية إذا كانت $\hat{G} < 0$

3. دالة الإنحدار اللوجستية للتمييز Logistic Regression Discriminant Function

تستخدم في حالة كون توزيع البيانات غير التوزيع الطبيعي. أي عندما يكون توزيع المتغيرات التوضيحية من عائلة التوزيع الأسي. وتعتمد هذه الطريقة على الاحتمالات السابقة واللاحقة لمجتمع (Y_2, Y_1) وتكون دالة التمييز التالية:

$$Z = \ln \frac{p(x \mid Y_1)}{p(x \mid Y_2)}$$

أما قاعدة التصنيف للمشاهدة x سيكون:

- تعود للمجموعة الأولى Y_1 إذا كانت $Z > 0$
- تعود للمجموعة الثانية Y_2 إذا كانت $Z < 0$

بعض الطرق اللامعلمية

تستخدم عندما يكون التوزيع غير طبيعي أو تكون هناك قيم شاذة مثل:

1. طريقة الرتب
2. الطرق اللا معلمية باستخدام التقديرات الحصينة (طريقة HUBER)
3. طريقة الدمج مابين طريقة الرتب و طريقة HUBER

طريقة الرتب

تستخدم بغض النظرعن توزيع البيانات سواء كان طبيعي أوغير طبيعي ويلجأ اليها الباحث عند عدم توفر الفرضيات الخاصة بالدالة المميزة فيتم إستخدام تحويلات الرتب للبيانات الاصلية.أي إستبدال البيانات الاصلية برتبها وبعدها يتم تطبيق الطرق السابقة (الخطية أو التربيعية).

1. طريقة HUBER

يتم إستبدال البيانات الأصلية ببيانات مشذبة. ولغرض تشذيب البيانات يتم حساب المتوسطات ومصفوفة التباين- التباين المشترك حيث يتم إستبدال $\bar{X}$ و S باوزان جديدة تعتمد على إحصاءة مهالونوبيس (D_i) وحسب الآتي:

$$\bar{X}^r = \frac{\sum w_i x_i}{\sum w_i}$$

$$S^r = \frac{\sum w_i^2 (x_i - \bar{X}^r)(x_i - \bar{X}^r)'}{\sum w_i^2}$$

$$w_i = \begin{cases} \dfrac{2}{D_i} & if \ D_i > 2 \\ 1 & if \ D_i \le 2 \end{cases}$$

وبعد الحصول على البيانات المشذبة يتم تطبيق الطرق السابقة (الخطية، التربيعية).

2. طريقة الدمج مابين طريقة الرتب و طريقة HUBER

نلجأ إلى اجراء تحويلين على البيانات الأصلية:

أولاً: تشذيب البيانات وجعلها قريبة من التوزيع الطبيعي بإستخدام أسلوب HUBER بالطريقة السابقة.

ثانياً: أخذ الرتب بالطريقة السابقة للبيانات المشذبة وبعدها تؤخذ البيانات الجديدة وتعالج على الطرق السابقة (الخطية، التربيعية).

الجانب التطبيقي

مثال1:

سُحبت عينة عشوائية مؤلفة من 12 رياضي من مجتمع طبيعي وأجرى عليهم اختباري اللياقة والكفاءة وذلك لغرض تصنيفهم إلى مهرة أوغير مهرة وكانت النتائج كما في الجدول أدناه حيث أن:

X_1 تمثل إختبار اللياقة و X_2 تمثل إختبار الكفاءة:

المهرة		غير المهرة	
X_2	X_1	X_2	X_1
33	60	35	57
36	61	36	59
35	64	38	59
38	63	39	61
40	65	41	63
		43	65
		41	59

الحل:

يجب أن نتحقق من شرطي الدالة التميزية الخطية:

1. البيانات تتوزع توزيعاً طبيعياً
2. مصفوفة التباين - التباين المشترك متساوية للمجتمعين.
   البيانات تم أخذها من المجتمع الطبيعي.
3. ومن ثم سوف نقوم باجراء إختبار لإثبات تساوي مصفوفة التباين - التباين المشترك للمجتمعين من خلال الفرضية

$$H_0 : \Sigma_1 = \Sigma_2$$
$$H_1 : \Sigma_1 \neq \Sigma_2$$

وأن إحصاءة الإختبار لهذه الفرضية Q وهي:

$$Q = MC^* \sim \chi^2_{(k-1)(p-1)}$$

حيث أن:

$$M = \sum_{i=1}^{2} n_i \ln|S| - \sum_{i=1}^{2} n_i \ln|S_i|$$

$$C^* = 1 - \frac{2p^2+3p-1}{6(p+1)(k-1)}\left[\sum \frac{1}{n_i} - \frac{1}{\sum n_i}\right]$$

$$S = \begin{pmatrix} 7.92 & 5.68 \\ 5.68 & 6.29 \end{pmatrix}$$

$$S_1 = \frac{\begin{pmatrix} S_{11} & S_{12} \\ & S_{22} \end{pmatrix}}{n_1 - 1} = \begin{pmatrix} 7.3 & 4.2 \\ 4.2 & 4.3 \end{pmatrix}$$

$$S_2 = \frac{\begin{pmatrix} S_{11} & S_{12} \\ & S_{22} \end{pmatrix}}{n_2 - 1} = \begin{pmatrix} 8.33 & 6.66 \\ 6.66 & 7.61 \end{pmatrix}$$

$|S| = 17.56$, $\ln|S| = 2.865$

$|S_1| = 13.7$, $\ln|S_1| = 2.62$

$|S_2| = 19.035$, $\ln|S_2| = 2.946$

$$M = 12(2.865) - [5(2.62) + 7(2.946)]$$
$$= 0.685$$

$$C^* = 1 - \frac{2(2)^2+3(2)-1}{6(2+1)(2-1)}\left[\left(\frac{1}{3}+\frac{1}{7}\right) - \frac{1}{12}\right]$$
$$= 0.8131$$

$$M\,C^* = (0.685)(0.8131) = 0.534$$

$$\chi^2_{1,\,0.05} = 3.841$$

ومن الواضح أن القيمة المحتسبة اقل بكثيرمن القيمة الجدولية:

$$H_0 : \Sigma_1 = \Sigma_2$$

لذا فالقرار هو عدم رفض (قبول) الفرضية، ومعنى ذلك عدم وجود فرق معنوي بين Σ_1 و Σ_2.

نلاحظ بأن شرطي الدالة التميزية الخطية متحقق. ولإيجاد دالة التميز الخطية وهي:

$$y = \underline{x}' S^{-1}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)$$

علينا أن نحسب القيم $\underline{\bar{x}_1}$ و $\underline{\bar{x}_2}$ و S لكل مجموعة واحتساب النتائج بموجبها وكما يلي:

المجموعة الأولى:

$$\sum X_1 = 182 \quad , \quad \sum X_1^2 = 6654$$
$$\sum X_2 = 313 \quad , \quad \sum X_2^2 = 19611$$
$$\sum X_1 X_2 = 11410$$
$$\underline{\bar{X}_1} = \begin{pmatrix} 36.4 \\ 62.6 \end{pmatrix}$$
$$S_{11} = \sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n} = 6654 - \frac{(182)^2}{5} = 29.2$$
$$S_{22} = \sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n} = 19611 - \frac{(313)^2}{5} = 17.2$$
$$S_{12} = \sum x_1 x_2 - \frac{\left(\sum x_1\right)\left(\sum x_2\right)}{n} = 11410 - \frac{(182)(313)}{5} = 16.8$$

المجموعة الثانية:

$$\sum X_1 = 273 \quad , \quad \sum X_1^2 = 10697$$
$$\sum X_2 = 423 \quad , \quad \sum X_2^2 = 25607$$
$$\sum X_1 X_2 = 16537$$
$$\underline{\bar{X}_2} = \begin{pmatrix} 39 \\ 60.42 \end{pmatrix}$$
$$S_{11} = \sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n} = 10697 - \frac{(273)^2}{7} = 50$$
$$S_{22} = \sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n} = 25607 - \frac{(423)^2}{7} = 45.71$$
$$S_{12} = \sum x_1 x_2 - \frac{\left(\sum x_1\right)\left(\sum x_2\right)}{n} = 16537 - \frac{(273)(423)}{7} = 40$$

ومن هذه النتائج نستخرج القيم التالية والتي تمثل التباينات المدمجة:

$$V_{11} = \frac{S_{11}(1) + S_{11}(2)}{n_1 + n_2 - 2} = \frac{29.2 + 50}{5 + 7 - 2} = 7.92$$
$$V_{22} = \frac{S_{22}(1) + S_{22}(2)}{n_1 + n_2 - 2} = \frac{17.2 + 45.71}{5 + 7 - 2} = 6.291$$
$$V_{12} = \frac{S_{12}(1) + S_{12}(2)}{n_1 + n_2 - 2} = \frac{16.8 + 40}{5 + 7 - 2} = 5.68$$

وبالتالي فإن مصفوفة التباين الكلي تكون:

$$|S| = 17.5623 \quad ,$$

$$S^{-1} = \frac{adj(S)}{|S|} = \frac{\begin{pmatrix} 6.291 & -5.68 \\ -5.68 & 7.92 \end{pmatrix}}{17.5623} = \begin{pmatrix} 0.358 & -0.323 \\ -0.323 & 0.451 \end{pmatrix}$$

الدالة المميزة

$$y = \underline{x}' S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2}) = \underline{x}' C^*$$

$$C^* = S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2})$$

$$= \begin{pmatrix} 0.358 & -0.323 \\ -0.323 & 0.451 \end{pmatrix} \begin{pmatrix} -2.6 \\ 2.18 \end{pmatrix} = \begin{pmatrix} -1.634 \\ 1.822 \end{pmatrix} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$$

$$y = (x_1 \quad x_2) \begin{pmatrix} -1.634 \\ 1.822 \end{pmatrix}$$

$$y = -1.634 x_1 + 1.822 x_2$$

نقطة الفصل

$$\bar{y}_1 = \underline{\bar{x}_1}' S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2}) = (36.4 \quad 62.6) \begin{pmatrix} -1.634 \\ 1.822 \end{pmatrix} = 54.5796$$

$$\bar{y}_2 = \underline{\bar{x}_2}' S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2}) = (39 \quad 60.42) \begin{pmatrix} -1.634 \\ 1.822 \end{pmatrix} = 46.359$$

$$z = \frac{\bar{y}_1 + \bar{y}_2}{2} = \frac{54.5796 + 46.359}{2} = 50.469$$

قاعدة التصنيف

المشاهدة x تعود للمجتمع الأول إذا كانت $y - z > 0$

المشاهدة x تعود للمجتمع الثاني إذا كانت $y - z \leq 0$

فلو افترضنا المشاهدة $\underline{X} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ لنرى لأي مجتمع تعود، وهنا يجب أن نجد قيمة الدالة المميزة:

$$y = -1.634 x_1 + 1.822 x_2$$
$$= -1.634(1) + 1.822(1) = 0.188$$

$$y - z = 0.188 - 50.469 = -50.281 < 0$$

إذا هذه المشاهدة يتم تصنيفها بأنها تعود للمجتمع الثاني.

أهمية كل متغير

بالإمكان إظهار أهمية كل متغير من خلال تطبيق المقياس الآتي:

$$C_i^* = C_i \sqrt{V_{ii}}$$
$$C_1^* = C_1 \sqrt{V_{11}} = -1.634\sqrt{7.92} = -4.598$$
$$C_2^* = C_2 \sqrt{V_{22}} = 1.822\sqrt{6.291} = 4.569$$

ونرى بأن المتغيرين لهما نفس الأهمية إذ الفرق قليل جداً.

إيجاد خطأ التصنيف

$$D^2 = (\underline{\bar{x}_1} - \underline{\bar{x}_2})' S^{-1} (\underline{\bar{x}_1} - \underline{\bar{x}_2})$$
$$= (-2.6 \quad 2.18) \begin{pmatrix} 0.358 & -0.323 \\ -0.323 & 0.451 \end{pmatrix} \begin{pmatrix} -2.6 \\ 2.18 \end{pmatrix} = 8.22$$
$$2.867 = D$$

وبذلك فإن إحتمال خطأ التصنيف أعلاه هو صغير جداً ويعطي إنطباعاً عن كفاءة عالية لدالة التمييز.

وللوقوف على مستوى هذه الكفاءة من جانب إستنتاجي إحصائي، نقوم بإختبار الفرضية التالية:

$$H_0 : \underline{\mu_1} = \underline{\mu_2}$$
$$H_1 : \underline{\mu_1} \neq \underline{\mu_2}$$

ويتم الإختبا بإستخدام إختبار F الذي يعتمد على إحصاءة هوتلنك Hotteling T^2 والتي تكون:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2 = \frac{(5)(7)}{5+7}(8.22) = 23.975$$

وبالتالي، فإن إختبار F سيكون:

$$F = \frac{n_1 + n_2 - p - 1}{n_1 + n_2 - 2} T^2$$
$$= \frac{5+7-2-1}{5+7-2}(23.975) = 21.577$$

وبالمقارنة مع القيمة الجدولية $f_{2,9,0.05} = 4.26$ فإننا نرفض H_0 مستنتجين فرق معنوي كبير بين الوسط الحسابي للمجتمعين مما يتيح القول بمستوى كفاءة عالي لدالة التمييز أعلاه.

مثال2:

سُحبت عينة عشوائية من 30 طالب لإجراء دراسة باستخدام التحليل المميز للتمييز بين الطلاب للصفوف الثلاث (المجاميع) C_1 , C_2 , C_3 للمرحلة الرابعة لقسم الأحصاء و ذلك بالإستناد على أربعة إختبارات وهي إختبارات متعدد (X_1)، الإستدلال (X_2)، العمليات (X_3)، التصميم (X_4). وكانت النتائج كما في الجدول أدناه والتي تمثل الأوساط الحسابية:

متغير الإختبار	C_1	C_2	C_3
X_1	64.5	60.5	58
X_2	86.4	81.8	83.1
X_3	75.2	73.4	71.4
X_4	81.9	88	86.6
حجم العينة	8	12	10

وأن مصفوفة التباين - التباين المشترك والمقدرة من العينة كانت كالاتي:

$$S = \begin{bmatrix} 1.00 & 0.5849 & 0.1774 & 0.1974 \\ 0.5849 & 1.00 & 0.2094 & 0.2170 \\ 0.1774 & 0.2094 & 1.00 & 0.2910 \\ 0.1974 & 0.2170 & 0.2910 & 1.00 \end{bmatrix}$$

الدالة المميزة

لدينا ثلاثة مجاميع لذلك توجد لدينا ثلاث دوال مميزة وهي y_{12} , y_{13} , y_{23} وكما يلي:

$$y_{12} = \underline{x}' S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2}) = \underline{x}' C_1^{*}$$

$$C_1^{*} = S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2})$$

$$|S| = 0.55406$$

$$S^{-1} = \frac{adj(S)}{|S|}$$

$$= \frac{1}{0.55406}\begin{pmatrix} 0.8508 & -0.4786 & -0.03503 & -0.0539 \\ -0.4786 & 0.86527 & -0.07559 & -0.0713 \\ -0.03503 & -0.07559 & 0.62194 & -0.15768 \\ -0.0539 & -0.0713 & -0.15768 & 0.62603 \end{pmatrix}$$

$$= \begin{pmatrix} 1.53557 & -0.8638 & -0.06322 & -0.09728 \\ -0.8638 & 1.56169 & -0.13642 & -0.128686 \\ -0.06322 & -0.13642 & 1.12251 & -0.2845 \\ -0.09728 & -0.128686 & -0.2845 & 1.12989 \end{pmatrix}$$

$C_1^* = S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_2})$

$$= \begin{pmatrix} 1.53557 & -0.8638 & -0.06322 & -0.09728 \\ -0.8638 & 1.56169 & -0.13642 & -0.128686 \\ -0.06322 & -0.13642 & 1.12251 & -0.2845 \\ -0.09728 & -0.128686 & -0.2845 & 1.12989 \end{pmatrix}\begin{pmatrix} 4 \\ 4.6 \\ 1.8 \\ -6.1 \end{pmatrix}$$

$$= \begin{pmatrix} 2.6484 \\ 4.26755 \\ 2.8755 \\ -8.3855 \end{pmatrix}$$

$$y_{12} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \end{pmatrix}\begin{pmatrix} 2.6484 \\ 4.26755 \\ 2.8755 \\ -8.3855 \end{pmatrix}$$

$= 2.6484\, x_1 + 4.26755\, x_2 + 2.8755\, x_3 - 8.3855\, x_4$

$y_{13} = \underline{x}' S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_3}) = \underline{x}' C_2^*$

$C_2^* = S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_3})$

$$= \begin{pmatrix} 1.53557 & -0.8638 & -0.06322 & -0.09728 \\ -0.8638 & 1.56169 & -0.13642 & -0.128686 \\ -0.06322 & -0.13642 & 1.12251 & -0.2845 \\ -0.09728 & -0.128686 & -0.2845 & 1.12989 \end{pmatrix}\begin{pmatrix} 6.5 \\ 3.3 \\ 3.8 \\ -4.7 \end{pmatrix}$$

$$= \begin{pmatrix} 7.3476 \\ -0.3746 \\ 4.7415 \\ -7.4485 \end{pmatrix}$$

$$y_{13} = (\,x_1 \quad x_2 \quad x_3 \quad x_4\,)\begin{pmatrix} 7.3476 \\ -0.3746 \\ 4.7415 \\ -7.4485 \end{pmatrix}$$

$$= 7.3476\ x_1 - 0.3746\ x_2 + 4.7415\ x_3 - 7.4485\ x_4$$

$$y_{23} = \underline{x}' S^{-1}\left(\underline{\bar{x}_2} - \underline{\bar{x}_3}\right) = \underline{x}'\, C_3^{*}$$

$$C_3^{*} = S^{-1}\left(\underline{\bar{x}_2} - \underline{\bar{x}_3}\right)$$

$$= \begin{pmatrix} 1.53557 & -0.8638 & -0.06322 & -0.09728 \\ -0.8638 & 1.56169 & -0.13642 & -0.128686 \\ -0.06322 & -0.13642 & 1.12251 & -0.2845 \\ -0.09728 & -0.128686 & -0.2845 & 1.12989 \end{pmatrix}\begin{pmatrix} 2.5 \\ -1.3 \\ 2 \\ 1.4 \end{pmatrix}$$

$$= \begin{pmatrix} 4.699 \\ -4.462 \\ 1.866 \\ 0.9369 \end{pmatrix}$$

$$y_{23} = (\,x_1 \quad x_2 \quad x_3 \quad x_4\,)\begin{pmatrix} 4.699 \\ -4.462 \\ 1.866 \\ 0.9369 \end{pmatrix}$$

$$= 4.699\ x_1 - 4.462\ x_2 + 1.866\ x_3 + 0.9369\ x_4$$

نقطة الفصل

نقطة التميز بين المجموعة الأولى والثانية هي:

$$Z_{12} = \frac{\bar{y}_1 + \bar{y}_2}{2}$$

$$\bar{y}_1 = \underline{\bar{x}_1}' S^{-1}\left(\underline{\bar{x}_1} - \underline{\bar{x}_2}\right) = (64.5 \quad 86.4 \quad 75.2 \quad 81.9)\begin{pmatrix} 2.6484 \\ 4.2675 \\ 2.8755 \\ -8.3855 \end{pmatrix} = 68.99$$

$$\bar{y}_2 = \underline{\bar{x}_2}' S^{-1}\left(\underline{\bar{x}_1} - \underline{\bar{x}_2}\right) = (60.5 \quad 81.8 \quad 73.4 \quad 88)\begin{pmatrix} 2.6484 \\ 4.2675 \\ 2.8755 \\ -8.3855 \end{pmatrix} = -17.55$$

$$Z_{12} = \frac{68.99 - 17{,}55}{2} = 25.72$$

نقطة التميز بين المجموعة الأولى والمجموعة الثالثة:

$$Z_{13} = \frac{\bar{y}_1 + \bar{y}_3}{2}$$

$$\bar{y}_1 = \underline{\bar{x}_1}' S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_3})$$

$$= (64.5 \quad 86.4 \quad 75.2 \quad 81.9)\begin{pmatrix} 7.3476 \\ -0.3746 \\ 4.7415 \\ -7.4485 \end{pmatrix} = 188.08$$

$$\bar{y}_3 = \underline{\bar{x}_3}' S^{-1}(\underline{\bar{x}_1} - \underline{\bar{x}_3}) = (58 \quad 83.1 \quad 71.4 \quad 86.6)\begin{pmatrix} 7.3476 \\ -0.3746 \\ 4.7415 \\ -7.4485 \end{pmatrix} = 88.53$$

$$Z_{13} = \frac{188.08 + 88.53}{2} = 138.305$$

نقطة التميز بين المجموعة الثانية والثالثة:

$$Z_{23} = \frac{\bar{y}_2 + \bar{y}_3}{2}$$

$$\bar{y}_2 = \underline{\bar{x}_2}' S^{-1}(\underline{\bar{x}_2} - \underline{\bar{x}_3})$$

$$= (60.5 \quad 81.8 \quad 73.4 \quad 88)\begin{pmatrix} 4.699 \\ -4.462 \\ 1.866 \\ 0.9369 \end{pmatrix} \quad = (58 \quad 83.1 \quad 71.4 \quad 86.6)$$

$$\begin{pmatrix} 4.699 \\ -4.462 \\ 1.866 \\ 0.9369 \end{pmatrix} = 116.11$$

$$Z_{23} = \frac{138.70 + 116.11}{2} = 127.405$$

قاعدة التصنيف

لكي نجد كل مشاهدة تعود إلى أي مجتمع نجد:

$W_{12} = y_{12} - Z_{12}$

$W_{13} = y_{13} - Z_{13}$

$W_{23} = y_{23} - Z_{23}$

تقع ضمن المجموعة الأولى إذا كان:

$W_{13} > 0 \quad , \quad W_{12} > 0$

تقع ضمن المجموعة الثانية إذا كان:

$$W_{13} > W_{12} \quad , \quad W_{13} > 0$$

تقع ضمن المجموعة الثالثة إذا كان:

$$W_{12} > W_{13} \quad , \quad W_{13} > 0$$

وفي ضوء ذلك، نستطيع عمل جدول تصنيفي للمشاهدات لنقف على نسبة التصنيفات الصحيحة (أي تصنيف المشاهدة لمجموعتها الأصلية) والتصنيفات الخاطئة (أي تصنيف المشاهدة ضمن أي من المجاميع الأخرى غير مجموعتها الأصلية).

التحليل العنقودي

Cluster Analysis (CA)

التحليل العنقودي مشابه للتحليل المميز من ناحية كونه ىيستخدم لتصنيف مجموعة من الأفراد أو الوحدات التجريبية إلى مجاميع فرعية معرفة بشكلٍ محدد وبلا تقاطع. الفرق بينهما هو أن التحليل المميز يمكن استخدامه عندما يكون لدى الباحث عينات عشوائية وكل واحدة منها مسحوبة مسبقاً من إحدى المجاميع الفرعية المحددة. بينما التحليل العنقودي يتعامل مع مسألة التصنيف عندما تكون الوحدات التجريبية غير معروفة مسبقاً لأي مجموعة فرعية تنتمي في الأصل.

ولكي نفهم بداية فكرة التحليل العنقودي، لنفترض أن صاحب شركة تجارية لبيع البضائع والمستلزمات يمتلك بياناتٍ ما تم جمعها من المستهلكين الذين يبتاعون من شركته. وهذه البيانات قد تتضمن متغيرات عديدة عن المستهلك مثل: العمر، المستوى التعليمي، مستوى الدخل، الحالة الزواجية، الحالة الوظيفية، عدد الأطفال دون سن الخامسة، وعددهم ما بين 13 - 6، وعدد من هم 14 سنة فأكثر.

وصاحب الشركة هذه ربما يرغب في تصنيف هؤلاء المستهلكين إلى مجاميع متباينة مستخدماً بياناته هذه وذلك لغرض إعلان سياسة العرض لبضائع معينة إعتماداً على طبيعة هذه المجاميع والتي تسمى (العناقيد Clusters).

هذه المجاميع بطبيعة الحال، توحي بتجانس ما بين المنتمين لكل مجموعة مع فروقات واضحة ما بين المنتمين لأي مجموعتين مختلفتين. وبشكلٍ عام، فإن الإسلوب العنقودي من شأنه تصنيف وحدات العينة إلى مجاميع (عناقيد Clusters) غير معروفة مسبقاً.

ومن الجدير بالذكر هنا أنه يجب عدم الخلط ما بين التحليل العنقودي والتحليل التمييزي. فالعنقودي لا يعتمد على أية معطيات تصنيفية مسبقة، بينما التمييزي يعتمد على كون عدد المجاميع ونوعيتها معرفةً مسبقاً ويتم جمع البيانات بموجبها ثم بعد ذلك تكون المهمة أن نقف على صحة عائدية المشاهدة لمجموعتها من عدمه. وفي حالة العدم، نقف على لأي مجموعة هي أقرب.

وبالتالى، فإن إسلوب التحليل العنقودي هدفه تجميع (تصنيف) المشاهدات وفق مجاميع Clusters بحيث كل مجموعة تحتوي على مشاهدات متجانسة قدر الإمكان بالنسبة لمتغيرات التعنقد المستخدمة. ولكي نبدأ تطبيق هذا الإسلوب، علينا إتباع الخطوات التالية:

1) تحديد مقياس التشابه Measure of Similarity
2) تحديد نوع إسلوب العنقدة المستخدم
3) تقرير عدد العناقيد
4) تفسير النتائج

والمهم هنا في مقياس التشابه هو الحصول على نمط بناء مجاميع سهلة لبيانات واسعة بإعتماد مبدأ التقارب أو التشابه. ولتحديد هذا المقياس، نجد أن الإختيار يعتمد على نوعية مقياس المتغيرات (إسمي، رتبي، فئوي، نسبي) أو طبيعتها (متقطعة ، مستمرة ، ثنائية). ومن الممكن إستخدام أحد القايييس التالية:

1- قياسات المسافة Distance Measures.
2- معامل الترابط (العلاقة) Association Coefficient

وفيما يلي توضيح ذلك.

مقياس المسافة Distance Measures

وهذا هو ما نعبر عنه بمسافة مهالانوبس Mahalanobis Distance ما بين أية مشاهدتين (نقطتين) X_s , X_r والتي تكون:

$$d(X_r , X_s) = d_{rs} = \sqrt{(Xr - Xs)' \sum^{-1} (Xr - Xs)}$$

حيث أن $\sum$ تمثل مصفوفة التباين والتباين المشترك ويمكن إحلال تقدير مناسب لها.

ملاحظة:

في الغالب نستخدم مقياس المسافة لهذا الغرض Mahalanobis Distance وقد نطلق عليه أيضاً المسافة الإقليدية Euclidean Distance وبالتالي فإن المقياس للمتغيرات يجب أن يكون كمياً (نسبي أو فئوي). وفي أدناه مثالاً على ذلك.

مثال:

لنفترض المثال البسيط التالي لغرض توضيح كيفية إستخدام إسلوب مقياس المسافة من خلال عينة من 6 أشخاص والبيانات التي تم جمعها عنهم تتمثل بمتغيرين هما:

X_1 : مقدار الدخل السنوي بالدولار

X_2 : مستوى التعليم بالسنوات

جدول البيانات الأولية

مستوى التعليم(سنة)	مقدار الدخل(ألف دولار)	الشخص(المشاهدة)
5	5	1
6	6	2
14	15	3
15	16	4
20	25	5
19	30	6

وبإستخدام مقياس المسافة Euclidean Distance بالنسبة لهذه البيانات، نحصل على مصفوفة التشابه (التماثل) Similarity Matrix التالية :

المشاهدة	1	2	3	4	5	6
1	0.0	2	181	221	625	821
2	2	0.0	145	181	557	745
3	181	145	0.0	2	136	250
4	221	181	2	0.0	106	212
5	625	557	136	106	0.0	26
6	821	745	250	212	26	0.0

والسؤال هو كيف بالإمكان إستخدام هذه البيانات الأولية ومقاييس المسافة في تحديد العناقيد؟.هنالك طريقتين رئيسيتين لعمل التحليل العنقودي وهما:

1- الطريقة الهرمية Hierarchical clustering
2- الطريقة اللاهر مية Non-hierarchical clustering

وفيما يلي سنتطرق للطريقة الهرمية لكونها المفضلة عموماً.

الطريقة الهرمية Hierarchical clustering

من خلال ملاحظة جدول البيانات الأولية نجد أن المشاهدتين الأولى والثانية متقاربتين مثلما هما المشاهدتين الثالثة والرابعة. في البداية بالطبع لدينا عناقيد بعدد المشاهدات ويمكن أن نبدأ بأي زوج منهما بداية الأمر ولنفترض أننا نختار المشاهدتين الأولى والثانية كخطوة أولى. وبذلك يتم الدمج لهاتين المشاهدتين في عنقود واحد وبالتالي يصبح لدينا خمسة عناقيد وفقاً للبيانات الأولية والمسافات.

الخطوة التالية هو تكوين مصفوفة جديدة من مقياس المسافات وفقاً لهذه العناقيد الخمسة.

ولأن العنقود رقم (1) المتشكل من المشاهدتين الأولى والثانية يتضمن مشاهدتين، فإنه يتوجب علينا إستخدام بعض القواعد لتحديد المسافات ما بين العناقيد الجديدة في مثل هذه الحالة.

وهنالك عدد من القواعد شائعة الإستخدام ومنها:

1- الطريقة المركزية Centroid method
2- طريقة الربط الفردي Single Linkage method
3- طريقة الربط المتكامل Complete Linkage method
4- طريقة الربط المتوسطي Average Linkage method
5- طريقة وارد Ward's method

وفيما يلي عرض بسيط لأول طريقتين.

الطريقة المركزية

وفقاً لهذه الطريقة، فإن كل مجموعة(عنقود) Cluster تم استبدالها بمعدل المشاهدات لتلك المجموعة عوضاً عن القيم الأصلية. وعلى سبيل المثال، عندما ندمج المشاهدتين الأولى والثانية لتكوين العنقود الأول والذي يكون مركزاً وسطياً بينهما. بمعنى إننا نستخدم معدل قيم هاتين المشاهدتين. وهذا يعني أن العنقود الأول لديه معدل مستوى التعليم ما يساوي 5.5 yr = 5 +6)/2) . ونستمر هكذا لحين الوصول إلى العنقود الأخير والذي يمكن أن يضم جميع المشاهدات.

طريقة الربط الفردي

في هذه الطريقة، فإن المسافة ما بين عنقودين يتم تمثيلها بأقل مسافة ما بين جميع الأزواج المحتملة للمشاهدات في كلا العنقودين. وأحد الأمثلة على هذه الطريقة هو إسلوب /طريقة أقرب الجوار Nearest neighbor method والتي تتضمن الخطوات التالية:

أ- إبدأ مع N من العناقيد حيث كل عنقود يتضمن مشاهدة واحدة.
ب- إدمج أقرب نقطتين وفقاً لإحدى طرق مقاييس المسافة المعتمد.
ت- إعتمد التباعد dissimilarity ما بين هذا العنقود الجديد وأي نقطة أخرى بمثابة أصغر مسافة بين هاتين النقطتين في العنقود وهذه النقطة الأخرى.
ث- الإستمرارية بدمج العناقيد الأكثر قرباً لبعضهما، وبالتالي سينخفض عدد العناقيد واحداً مع كل خطوة. إن التباعد ما بين أي عنقودين هو دائماً المسافة ما بين أقرب عنقودين.

ولذلك فإن طريقة "أقرب الجوار" تبدأ أيضاً مع N من العناقيد حيث كل عنقود يتضمن مشاهدة واحدة، وتستمر بدمج النقاط والعناقيد حتى تنتهي العملية بعنقود واحد يتضمن جميع المشاهدات.

وإزاء ذلك، فإنه من الواضح أن العدد المناسب من العناقيد التي ننتهي عندها يقع ما بين عددها عند البداية وعددها عند النهاية. وهنالك بعض الطرق لتحديد مكان التوقف لعملية الدمج هذه بضمنها النظرة المنطقية في هذا الشأن.

إحدى هذه الطرق التي تساعد في وقف عملية الدمج هي من خلال بناء شكل الشجرة الهرمي. وهذا الشكل يتضمن فروعاً تربط نقاط البيانات وتعكس بالترتيب الذي تإخذه النقاط لضمها للعناقيد. وطول فروعها يجب أن تتناسب مع المسافات ما بين النقاط والعناقيد عندما يتم دمج النقاط بالعناقيد. وسيتم توضيح هذه الآلية من خلال المثال التالي.

مثال:

لنفترض مصفوفة المسافات الإقليدية (مصفوفة التباعد) ما بين 6 من المشاهدات. وأعطينا هذا المثال بهذه القيم لتسهيل فهم عملية الإجراءات المتخذة وفقاً لما تم تبيانه في أعلاه:

العنقود (المشاهدة)	1	2	3	4	5	6
1	0.0	0.31	0.23	0.32	0.26	0.25
2		0.0	0.34	0.21	0.36	0.28
3			0.0	0.31	(0.04)	0.07
4				0.0	0.31	0.28
5					0.0	0.09
6						0.0

وقبل البدء برسم الشكل الشجري الهرمي، دعنا نتعامل في عنقدة هذه البيانات وفقاً لطريقة الربط الفردي وحسب الخطوات التالية:

في البداية، وكما قلنا، نعتبر عدد العناقيد مساوياً لعدد المشاهدات (النقاط). فإذا افترضنا الرمز C لمجموعة العناقيد، ففي هذه الحالة نبدأ بالمجموعة C_0 وتكون الآتي:

$C_0 = \{ [1] , [2] , [3] , [4] , [5] , [6] \}$

أي أننا لدينا الآن 6 عناقيد. وفي ضوء ذلك تكون خطوات العمل التالية كما يلي:

1- بملاحظة مصفوفة المسافات أعلاه، نجد أن النقطتين الأقرب لبعضهما هما (3) و (5) حيث أن المسافة هي (0.04) وبذلك فإن الخطوة الأولى هي دمج هاتين النقطتين في عنقود واحد وبذلك تصبح لدينا مجموعة العناقيد C_1 وهي

$C_1 = \{ [1] , [2] , [3 , 5] , [4] , [6] \}$

وبذلك يترتب علينا صياغة مصفوفة مسافات جديدة وفقاً إلى C_1 وهي الآتي:

العنقود	[1]	[2]	[3 , 5]	[4]	[6]
[1]	0.0	0.31	0.23	0.32	0.25
[2]		0.0	0.34	0.21	0.28
[3 , 5]			0.0	0.31	(0.07)
[4]				0.0	0.28
[6]					0.0

ملاحظة:

إن المصفوفة الجديدة هي عبارة عن المصفوفة السابقة بعد حذف العمود والصف للعنقود المدمج مع سابقه. ولأننا قمنا بدمج العنقود [5] مع العنقود [3] فإن المصفوفة أعلاه عبارة عن المصفوفة الأصلية محذوفاً منها عمود وصف [5].

وهذه المسافات محسوبة على أساس مقارنة النقطة [1] مع [5 , 3] ونختار الأصغر من بين 0.23 و 0.26 فوضعنا 0.23 الأصغر. وعلى نفس المنوال، تم تحديد المسافات الجديدة ما بين بقية النقاط (العناقيد) جميعاً لنرى الوضع أعلاه.

2- ومن ملاحظة ذلك، نجد أن أقرب مسافة هي ما بين [6] و [5 , 3] وهي (0.07) وبذلك يتم دمج هذين العنقودين لتكون لدينا المجموعة C_2 التالية:

C_2 = { [1] , [2] , [3 , 5 , 6] , [4] }

وفي ضوء ذلك يترتب علينا صياغة مصفوفة مسافات جديدة والتي يمكن تحديدها من خلال الخطوة السابقة وهذا هو الجانب الإيجابي في طريقة الربط المنفرد والتي تغنينا عن الرجوع للمصفوفة الأولية. وعلى هذا الأساس، وفي ضوء C_2 ستكون المصفوفة الجديدة للمسافات كما يلي:

العنقود	[1]	[2]	[3 , 5 , 6]	[4]
[1]		0.31	0.23	0.32
[2]			0.28	(0.21)
[3 , 5 ,6]				0.28
[4]				

3- وهذه المصفوفة تنتج لنا تقارب [2] مع [4] حيث المسافة الأقل وهي (0.21) وبذلك تكون مجموعة العناقيد الجديدة C_3 بالشكل التالي:

C_3 = { [1] , [2 , 4] , [3 , 5 , 6] }

وفي ضوء C_3 هذه ستكون المصفوفة الجديدة للمسافات كما يلي:

العنقود	[1]	[2 , 4]	[3 , 5 , 6]
[1]	0.0	0.31	(0.23)
[2 , 4]		0.0	0.28
[3 , 5 ,6]			0.0

4- وهذه المصفوفة تشير إلى أقرب مسافة ما بين [1] و [6 , 5 , 3] وهي (0.23). ولذلك سيتم الدمج بينهما لتصبح مجموعة العناقيد الجديدة C_4 بالشكل التالي:

C_4 = { [3 , 5 , 6 , 1] , [2 , 4] }

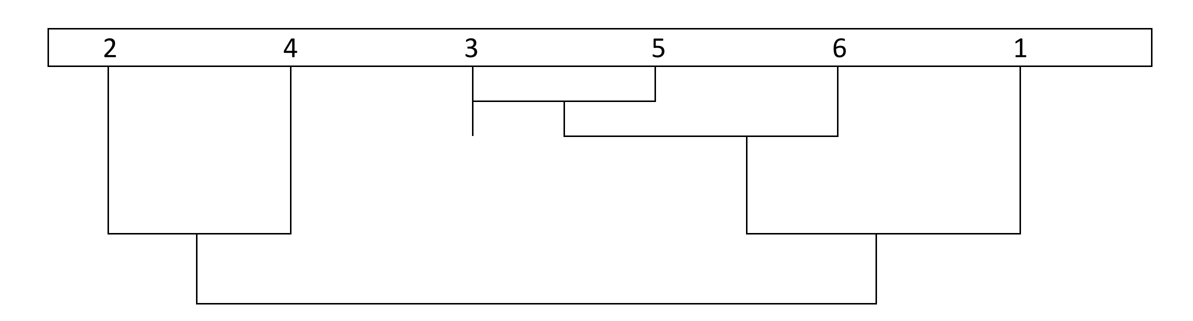
وأن مصفوفة المسافات في ضوء C_4 ستكون:

العنقود	[3 , 5 ,6 ,1]	[2 , 4]
[3 , 5 ,6 , 1]	0.0	0.28
[2 , 4]		0.0

5- إن عملية الدمج الأخيرة هي بالتالي ما بين هاتين المجموعتين فتكون لدينا مجموعة (عنقود) واحد وهي:

C_5 = ([1 , 2 , 3 , 4 , 5 , 6] }

وفي ضوء هذه النتائج يمكننا تنفيذ الشكل البياني للشجرة الهرمية وعلى النحو المبين في أدناه:

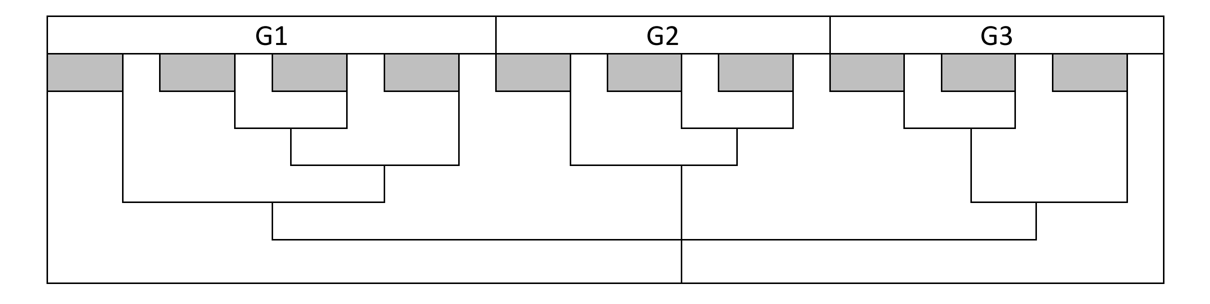


وهذا الشكل يساعد في إتخاذ قرار التوقف حيث قد نرى التوقف عند العنقودين

[2 , 4] و [3 , 5 ,6 , 1]

أو ما قبل ذلك وعند العناقيد الثلاثة [1] و [2 , 4] و [3 , 5 , 6]

ولكي يكون الموضوع أكثر وضوحاً حول إختيار توقف العنقدة، لنفترض أن حصيلتنا بالشكل التالي:



وهذا يمثل الشكل البياني المثالي الوضوح للشجرة الهرمية. وفي ضوء ذلك فقد نختار الإبقاء على ثلاثة عناقيد كإختيار معقول وهي G_1 و G_2 و G_3

وبالعودة لمثالنا الأول حيث البداية بالمصفوفة:

المشاهدة	1	2	3	4	5	6
1	0.0	(2)	181	221	625	821
2		0.0	145	181	557	745
3			0.0	2	136	250
4				0.0	106	212
5					0.0	26
6						0.0

وعند تطبيق طريقة الربط الفردي، تكون لدينا النتائج التالية:

1- الخطوة الأولى دمج [1] مع [2] فتكون المجاميع

C_1 = { [1 , 2] , [3] , [4] , [5] , [6]

2- وبإعتماد المصفوفة الجديدة للمسافات وهي:

العنقود	[1 , 2]	[3]	[4]	[5]	[6]
[1 , 2]	0.0	181	221	625	821
[3]		0.0	(2)	136	250
[4]			0.0	106	212
[5]				0.0	26
[6]					0.0

ومنها نجد إدماج [3] و [4] لتصبح لدينا المجموعة:

C_2 = { [1 , 2] , [3 , 4] , [5] , [6] }

3- وبإعتماد ذلك، تكون لدينا المصفوفة الجديدة:

العنقود	[1 , 2]	[3 ,4]	[5]	[6]
[1 , 2]	0.0	181	625	821
[3 , 4]		0.0	136	250
[5]			0.0	(26)
[6]				0.0

4- ومنها نجد إدماج [5] مع [6] لتصبح لدينا المجموعة:

C_3 = { [1 , 2] , [3 , 4] , [5 ,6] }

وبذلك تنتج لدينا مصفوفة مسافات جديدة وهي:

العنقود	[1 , 2]	[3 ,4]	[5 , 6]
[1 , 2]	0.0	181	625
[3 , 4]		0.0	(136)
[5 ,6]			0.0

ومنها نجد أن [5 , 6] أقرب إلى [3 ,4] فيتم الدمج بينهما لتصبح لدينا المجموعة:

C_4 = { [1 , 2] , [3 , 4 , 5 ,6] }

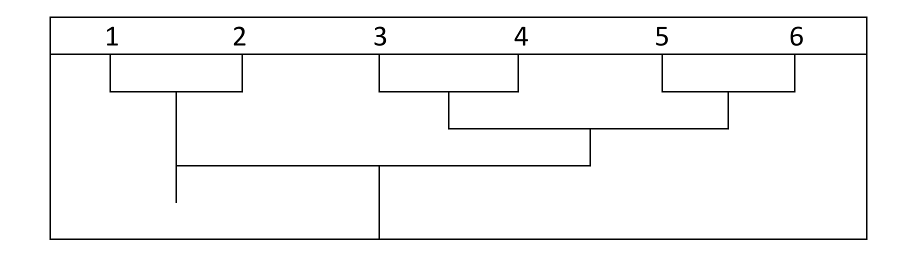
والنتيجة النهائية لمصفوفة المسافات تكون :

العنقود	[1 , 2]	[3 ,4 , 5 ,6]
[1 , 2]	0.0	181
[3 , 4 ,5 , 6]		0.0

5- وأخيراً ليس لدينا سوى دمج هاتين المجموعتين (العنقودين) لتصبح لدينا المجموعة الأخيرة وبواقع عنقود واحد وهو:

C_4 = { [1 , 2 , 3 , 4 , 5 , 6] }

والشكل الشجري الهرمي لهذه النتيجة يكون:



ونرى أن نتوقف عند العناقيد:

C_3 = { [1 , 2] , [3 , 4] , [5 ,6] }

وفيما يلي مثالاً لبيانات حقيقية لواقع الجرائم المختلفة (القتل، الإغتصاب، السرقة ،.......، سرقة السيارات) المسجلة في 6 مدن أمريكية (أطلنطا، بوسطن، شيكاغو، دالاس، دنفر، ديترويت) محسوبة لكل 100,000 من السكان حيث تم إحتساب مصفوفة المسافات ما بين هذه المدن وفقاً لبيانات الجرائم فكانت حسب الآتي[3]:

العنقود (المدن)	1 أطلنطا	2 بوسطن	3 شيكاغو	4 دالاس	5 دنفر	6 ديترويت
أطلنطا 1	0	536	516	590	693	716
بوسطن 2	536	0	447	833	915	881
شيكاغو 3	516	447	0	924	1073	971
دالاس 4	590	833	924	0	527	464
دنفر 5	693	915	1073	527	0	(358)
ديترويت 6	716	881	971	464	(358)	0

أصغر قيمة للمسافات محصورة بين قوسين ( )

ولكون أصغر مسافة في الجدول أعلاه هي (358) وهي المسافة ما بين المدينتين دنفر (رقم 5) و دترويت (رقم 6)، فإننا ندمج هاتين المدينتين بإعتبارهما يشكلان عنقوداً واحداً من حيث سجل الجرائم المرتكبة. وبعد عملية الدمج بين هاتين المدينتين وما تبع ذلك من الحذف والتعويض، أصبح لدينا الجدول التالي:

العنقود (المدن)	1	2	3	4	[5 , 6]
1	0.0	536	516	590	693
2	536	0.0	(447)	833	881
3	516	(447)	0.0	924	971
4	590	833	924	0.0	464
[5 , 6]	693	881	971	464	0.0

ولكون أصغر مسافة في الجدول أعلاه هي (447) وهي المسافة ما بين المدينتين بوسطن (رقم 2) وشيكاغو (رقم 3)، فإننا ندمج هاتين المدينتين بإعتبارهما يشكلان عنقوداً واحداً من حيث سجل الجرائم المرتكبة. وبعد عملية الدمج بين هاتين المدينتين وما تبع ذلك من الحذف والتعويض، أصبح لدينا الجدول التالي:

العنقود (المدن)	1	[2 , 3]	4	[5 , 6]
1	0.0	516	590	693
[2 , 3]	516	0.0	833	881
4	590	833	0.0	(464)
[5 , 6]	693	881	(464)	0.0

وبنفس الطريقة تدمج مدينة دالاس (رقم 4) مع المدينتين دنفر (رقم 5) وديترويت (رقم 6) ليصبح لدينا الجدول التالي:

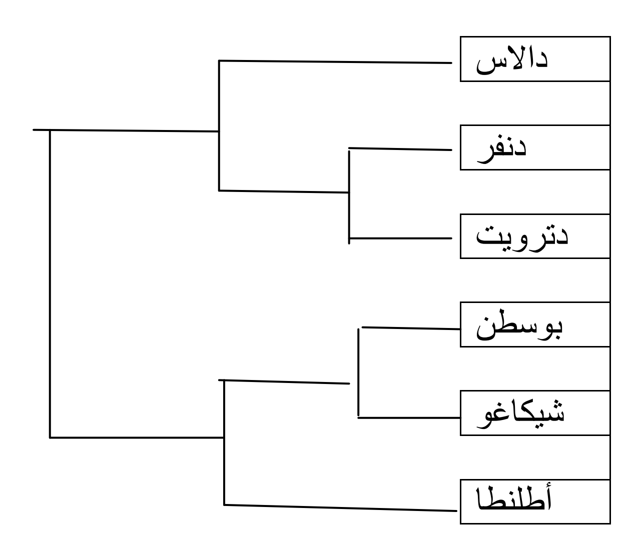
العنقود (المدن)	1	[2 , 3]	[4 ,5 , 6]
1	0.0	(516)	590
[2 , 3]	(516)	0.0	833
[4 , 5 , 6]	590	833	0.0

وهنا يتم دمج مدينة أطلنطا (رقم 1) مع المدينتين بوسطن (رقم 2) وشيكاغو (رقم 3) ليصبح لدينا الجدول التالي :

العنقود (المدن)	[1 , 2 , 3]	[4 , 5 , 6]
[1 , 2 , 3]	0.0	590
[4 , 5 , 6]	590	0.0

والذي أصبح لدينا بموجبه عنقودين (أطلنطا، بوسطن، شيكاغو) و (دالاس، دنفر، دترويت).

وخلاصة التحليل تشير إلى أن دنفر أكثر قرباً إلى دترويت ومن بعد ذلك كانت بوسطن أكثر قرباً إلى شيكاغو. وبعد الدمج وجدنا دالاس تقترب لمجموعة دنفر/ دترويت ثم بعد ذلك نجد أطلنتا تقترب لمجموعة بوسطن/ شيكاغو. والشكل التالي يوضح ذلك:



تحليل الإرتباط القويم (إرتباط المجموعات)
Canonical Correlation Analysis

إن تحليل الإرتباط القويم هو بمثابة تعميم للإرتباط المتعدد Multiple Correlation في مسائل الإنحدار المتعدد.ففي الإرتباط المتعدد يكون لدينا متغير معتمد أحادي Y يرتبط بمتغيرين توضيحيين أو أكثر (X_1 , X_2 , ... , X_p) ليتضح لدينا مدى علاقته بها. أما في الإرتباط القويم، فيكون لدينا عدد q من المتغيرات المعتمدة (Y_1 , Y_2 , ... , Y_q) بدلاً من الواحد ونريد معرفة مدى وشكل إرتباطاتها مع مجموعة المتغيرات (X_1 , X_2 , ..., X_p). أي أننا نعني بالإرتباط القويم هنا بأنه الإرتباط بين مجموعتين من المتغيرات إحداهما مجموعة (X_1 , X_2 , ..., X_p) والأخرى مجموعة (Y_1 , Y_2 , ... , Y_q) عن طريق إيجاد أكبر ترابط خطي للمتغيرات في إحدى المجموعتين مع التراكيب الخطية للمتغيرات في المجموعة الثانية والمجموعتين ذات توزيع مشترك. ووفقاً لذلك، فإن تطبيق هذا التحليل يتطلب تقسيم المتغيرات الناتجة إلى المجموعتين. هذا التقسيم يكون على أساس طبيعة المتغير وليس على أساس التحري والفحص للبيانات. مثال ذلك يمكن دراسة الإرتباط ما بين سمات معينة لدى الآباء وسمات أخرى لدى الأبناء للوقوف على مدى الترابط ما بين الأباء والأبناء فيما يخص مجال معين يتم جمع البيانات حولها من خلال متغيرات موصوفة مسبقا للمجتمعين. كما أن القياس المناسب لكلا المجموعتين من المتغيرات هو النسبي أو الفئوي فقط.

في بعض الأحيان قد يجد الباحث صعوبة في جمع بيانات لمتغيرات محددة تغطي مجال معين ولكنه بحاجة لدراسة هذا المجال. هذه الصعوبة قد تكون نتيجة عدم توفر البيانات أو كلفتها العالية. ما الحل إذا ونحن نعرف أن علم الإحصاء لا يقف عاجزاً أمام مثل هذه الحالات وغيرها، على الباحث هنا اللجوء إلى ما يسمى بالمتغيرات المساعدة Auxiliary Variables لتكون بديلاً عن المتغيرات الأصلية. والتأريخ يزخر بأمثلة من هذا الإجراء. والتأريخ الإسلامي يذكر أن معركة كانت على وشك أن تدور بين المسلمين وجيش الروم وأراد القائد معرفة تعداد جيش الروم بشكلٍ غير مباشر فلجأ إلى معرفة ذلك تقديراً من خلال عدد أرغفة الخبز أو الذبائح التي يستهلكها ذلك الجيش والذى كان ممكناً معرفتها إلى حدٍ ما.

إن تحليل الإرتباط القويم قد يساهم في هذا الجانب من خلال توصيف مجموعة متغيرات أخرى كمتغيرات مساعدة مقابلة للمتغيرات الأصلية وجمع البيانات عنها ممكناً. إن درجة الثقة بإمكانية إعتماد هذه المتغيرات المساعدة بديلاً عن المتغيرات الأصلية في الدراسة تتناسب طردياً مع قيمة الإرتباط القويم الذي نحصل عليه بين مجموعتي المتغيرات.

أحد التساؤلات الأساسية التي على تحليل الإرتباط القويم الإجابة عنها هو ما إذا كان بالإمكان إستخدام المتغيرات في إحدى المجموعتين للتنبؤ بالمتغيرات في المجموعة الأخرى.

فعندما يكون ذلك ممكناً، فإن ذلك يعني بأن تحليل الإرتباط القويم يعمل على تلخيص العلاقات ما بين مجموعتي المتغيرات من خلال تكوين متغيرات جديدة من كلٍ من مجموعتي المتغيرات الأصلية.

ومن الجدير بالذكر أن مفهوم الإرتباط القويم ظهر في الفترة 1936/1935 من قبل الإحصائي هوتلنك Hotelling حيث ذكر أن تحليل الإرتباط المتعدد ما هو إلا حالة خاصة من الإرتباط القويم. وفي عام 1940 كان فيشر أول من إستخدم الإرتباط القويم لتحليل الجداول التوافقية ذات الإتجاهين (rxc) ذات فئات مرتبة.

وعند التطبيق هنا لو افترضنا وجود مجموعة متغيرات $(X_1 , X_2 , ... , X_p)$ ومجموعة متغيرات أخرى $(Y_1 , Y_2 , ... , Y_q)$ وأن $(q<p)$ فإنه ستكون لدينا حصيلة إرتباطات ثنائية بعدد $(p + q)$ ننطلق منها في عملية تحليل الإرتباط القويم. هذه العملية تشمل تحديد الإرتباط ما بين المتغيرات القويمة V_i و U_i حيث:

$$U_1 = a_{11} X_1 + a_{12} X_2 + \cdots\cdots + a_{1p} X_p$$

$$U_2 = a_{21} X_1 + a_{22} X_2 + \cdots\cdots + a_{2p} X_p$$

$$\vdots$$

$$U_r = a_{r1} X_1 + a_{r2} X_2 + \cdots\cdots + a_{rp} X_p$$

وكذلك:

$$V_1 = b_{11} Y_1 + b_{12} Y_2 + \cdots\cdots + b_{1q} Y_q$$

$$V_2 = b_{21} Y_1 + b_{22} Y_2 + \cdots\cdots + b_{2q} Y_q$$

$$\vdots$$

$$V_r = b_{r1} Y_1 + b_{r2} Y_2 + \cdots\cdots + b_{rq} Y_q$$

وأن (r) هنا يمثل أصغر الأعداد $(p$ و $q)$. وهذه العلاقات الخطية يتم تحديدها بحيث نحصل على أعلى إرتباط ما بين U_1 و V_1. كذلك يكون لدينا الإرتباط الأعلى التالي ما بين U_2 و V_2 وهكذا. بمعنى آخر فإن كل زوج من المتغيرات القويمة (U_1 , V_1) و (U_2 , V_2) و.......و (U_r , V_r) يمثل إتجاهاً مستقلاً في العلاقة ما بين مجموعتي المتغيرات $(X_1 , X_2 , ... , X_p)$ و $(Y_1 , Y_2 , ... , Y_q)$. ولأن أول زوج (U_1 , V_1) من المتغيرات القويمة يملك أعلى إرتباط، فإنه يعتبر الأكثر أهمية. وأن الزوج الثاني الذي يليه وهو (U_2 , V_2) له ثاني أعلى إرتباط، وبالتالي فإنه يليه في الأهمية ويكون ثاني أهم زوج وهكذا بالنسبة

لجميع أزواج المتغيرات القويمة إلى أن نأتي على الزوج الأخير وهو $(U_r\ ,\ V_r)$ والذى هو الأقل أهمية لكونه ذو الإرتباط الأصغر.

طرق تنفيذ تحليل الإرتباط القويم

لعله من السهل برمجة الحسابات المتعلقة بتحليل الإرتباط القويم لتنفيذها في الحاسبة شرط أن يكون البرنامج مناسب للتعامل مع المصفوفات الجبرية لأن الأساس في العملية الحسابية هو البدء بتحديد مصفوفات الإرتباط ما بين متغيرات كل مجموعة من المتغيرات على انفراد بالإضافة إلى مصفوفة الإرتباط ما بين المجموعتين. ونعني بذلك لو كانت لدينا مجموعتي المتغيرات $X_1\ ,\ X_2\ ,\ ....,\ X_p$ و $Y_1\ ,\ Y_2\ ,\ ....\ ,\ Y_q$ فإنه سيكون لدينا مصفوفة الإرتباطات التربيعية ما بين جميع هذه المتغيرات والتي ستكون بأبعاد (p+q)x(p+q) وفقاً لما يلي:

$$
\begin{array}{c c}
 & \begin{array}{cc} X_1\ \ X_2 \ldots X_p & Y_1\ \ Y_2\ \ldots.\ \ Y_q \end{array} \\
\begin{array}{c} X_1 \\ \vdots \\ X_p \\ \\ Y_1 \\ \vdots \\ Y_q \\ \\ \end{array} &
\left[\begin{array}{cc}
p\times p\,matrix & p\times q\,matrix \\
for\ \ X's\,Variables & for\ \ XY's\,Variables \\
A & C \\
 & \\
q\times p\,matrix & q\times q\,matrix \\
for\ YX's\,Variables & for\ Y's\,Variables \\
C' & B
\end{array}\right]
\end{array}
$$

ومن هذه المصفوفة الرئيسية يمكننا تكوين المصفوفة $B^{-1}C'A^{-1}C$ ذات الأبعاد $q\times q$ لغرض إستخدامها في تحديد القيم المميزة $\lambda_1 > \lambda_2 \ \ > \cdots\cdots > \lambda_r$ لكون (r=q<p) والتي تعتبر تربيعات لقيم الإرتباطات القويمة ما بين المتغيرات القويمة المقابلة لها. وهذا بطبيعة الحال يتم من خلال حل مجموعة المعادلات الناتجة عن محددة المصفوفة

$$\left|B^{-1}C'A^{-1}C-\lambda I\right|=0 \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

ولغرض استكمال الصورة التي تكون عليها المتغيرات القويمة، فإننا نحتاج إلى تحديد المتجهات المميزة $\mathbf{b_1\ ,\ b_2\ ,\ ....,\ b_r}$ والتي هي عبارة عن الأوزان القويمة لمجموعة المتغيرات Y's ضمن المتغيرات القويمة **V's** , وهذا يتم من خلال معادلة المصفوفات:

$$\left(B^{-1}C'A^{-1}C-\lambda I\right)\mathbf{b} = 0$$

والتي يمكن تحويلها للصيغة:

$$\left(C'A^{-1}C-\lambda B\right)\mathbf{b} = 0 \qquad .... (2)$$

أما المتجهات $\mathbf{a_1}$, $\mathbf{a_2}$, ... , $\mathbf{a_r}$ والتي هي عبارة عن الأوزان القويمة لمجموعة المتغيرات X's ضمن المتغيرات القويمة U's ، فيتم تحديدها من خلال معادلة المصفوفات:

$$\mathbf{a_i} = A^{-1}C\ \mathbf{b_i} \qquad ....(3)$$

ومن الجدير بالذكر أن هذه الحسابات في أعلاه تتم جميعها بإستخدام القيم المعيارية للمتغيرات الأصلية والتي تكون أوساطها صفراً وانحرافها المعياري وحدة واحدة.

ووفقاً لهذه الأوزان القويمة $\mathbf{a_i}$ و $\mathbf{b_i}$ تكون المتغيرات القويمة للمجموعتين على الشكل الآتي:

$$U_i = \mathbf{a'_i}\,\mathbf{X} = \left(a_{i1}, a_{i2}, ..., a_{ip}\right)\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_p \end{bmatrix} = a_{i1}x_1 + a_{i2}x_2 + ... + a_{ip}x_p \qquad , \ i=1,2,.....,r$$

$$V_i = \mathbf{b'_i}\,\mathbf{y} = \left(b_{i1}, b_{i2}, ..., b_{iq}\right)\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_q \end{bmatrix} = b_{i1}y_1 + b_{i2}y_2 + ... + b_{iq}y_q \qquad , \ i=1,2,.....,r$$

بطبيعة الحال، عندما نتكلم عن الإرتباط القويم ما بين زوج المتغيرات القويمة، فإنه يكون معلوماً لدينا أن الإرتباط بينهما ينعكس من عدد المشاهدات n وهو حجم العينة المستخدمة والتي تكون بالشكل التالي:

$$\begin{array}{cccccc} x_{11} & x_{21}.....x_{p1} & y_{11} & y_{21}......y_{q1} \\ x_{12} & x_{22}.....x_{p2} & y_{12} & y_{22}......y_{q2} \\ \vdots & \vdots \quad \vdots & \vdots & \vdots \quad \vdots \\ \vdots & \vdots \quad \vdots & \vdots & \vdots \quad \vdots \\ x_{1n} & x_{2n}.....x_{pn} & y_{1n} & y_{2n}......y_{qn} \end{array}$$

وحيث أنه سيكون لدينا العدد الكلي للإرتباطات القويمة (r) فإنه ليس من المنطقي أن نعتمدها جميعاً للتحليل وإنما سنعتمد عدداً محدوداً جداً وذلك وفقاً لمعنويتها في إختبارٍ مناسبٍ لها وكما هو في القسم التالي.

إختبار معنوية الإرتباط القويم

من البديهي أنه طالما أن مربع أي إرتباط قويم يساوي القيمة المميزة المقابلة، فهذا يعني أن الإرتباط القويم الأول يعتبر الأعلى ويليه الثاني وهكذا فإن الأخير هو الأصغر. لذلك لو إفترضنا أننا وجدنا الإرتباط القويم الثاني غير معنوي، فلا حاجة لنا بإجراء الإختبارات للثالث وما بعده. وطريقة الإختبار هي الإختبار التقريبي الذي تم إقتراحه من قبل بارتليت Bartlett(1947) لمعرفة عدد الإرتباط القويمة المعنوية وهي بسيطة وسهلة الإستخدام.

هذه الطريقة تعتمد بداية على إحصاءة الإختبار:

$$\Delta_0^2 = -\left\{n - \frac{1}{2}(p + q + 1)\right\}\sum_{i=1}^{r} \ln(1 - \lambda_i)$$

والتي تتبع توزيع مربع كاي بدرجات حرية مساوية إلى (pq) أي (χ_{pq}^2). فإذا كانت معنوية (أي أن $\Delta_0^2 > \chi_{pq}^2$)، فإنه يكون لدينا واحداً في الأقل من الإرتباطات القويمة وهو الأول معنوياً وهو ما يكون في الغالب.

ووفقاً لمعنوية الإرتباط القويم الأول، فإننا نستمر لفحص الإرتباط الذي يليه وفقا لنفس الصيغة ولكن بعد استبعاد ما يتعلق بأثر الإرتباط الأول من إحصاءة الإختباروفقاً لما يلي:

$$\Delta_1^2 = -\left\{n - \frac{1}{2}(p + q + 1)\right\}\sum_{i=2}^{r} \ln(1 - \lambda_i)$$

والتي تتبع توزيع مربع كاي بدرجات حرية مساوية إلى (p-1)(q-1) أي ($\chi_{(p-1)(q-1)}^2$). فإذا كانت معنوية (أي أن $\Delta_1^2 > \chi_{(p-1)(q-1)}^2$)، فإنه يكون لدينا الإرتباطات القويمة الأول والثاني معنويان. وهذا يدفعنا للتطبيق التالي لإحصاءة الإختبار بعد استبعاد ما يتعلق بأثر هذان الإرتباطان القويمان منها وهكذا حتى نصل إلى آخر إرتباط معنوي. وبشكلٍ عام فإننا لو وجدنا عدد (j) منها معنوياً فإننا نطبق الصيغة التالية :

$$\Delta_j^2 = -\left\{n - \frac{1}{2}(p + q + 1)\right\}\sum_{i=j+1}^{r} \ln(1 - \lambda_i)$$

والتي تتبع توزيع مربع كاي بدرجات حرية مساوية إلى ((p-j)(q-j) أي ($\chi_{(p-j)(q-j)}^2$).

ومن الجدير بالذكر أنه لو حصل لدينا عدم معنوية في أي خطوة مما ذكرنا في أعلاه، فإنه يجدر بنا التوقف وإعتبار الإرتباطات القويمة التالية أيضاً ليست معنوية. كما نلاحظ أن حجم العينة (n) هنا له أثر واضح في قيمة دالة الإختبار وبالتالي في معنويته.

تحليل نتائج المتغيرات القويمة

بعد الوقوف على نتائج الإختبارات أعلاه وحصولنا على نتائج معنوية، فإن الخطوة التالية هو الخروج بنتائج تحليلية لهذه النتائج ودلالة القياسات التي تعكسها. وبالتأكيد فإن ذلك يشمل جميع القيم المعنوية للإرتباطات القويمة ولو أن التركيز سيكون على الأول وهو الأكبر وقد نكتفي بذلك حتى في حالة كون قيمته غير معنوية.

فإذا إعتبرنا المتغيرات القويمة:

$$U_i = a_{i1} X_1 + a_{i2} X_2 + \cdots\cdots + a_{ip} X_p$$

و

$$V_i = b_{i1} Y_1 + b_{i2} Y_2 + \cdots\cdots + b_{iq} Y_q$$

فإنه من الممكن توضيح U_i بالنسبة إلى أي من مجموعة المتغيرات $X_1 , X_2 , ... , X_p$ التي لها أكبر أوزان قويمة a_{ij} وكذلك فيما يخص V_i بالنسبة إلى مجموعة المتغيرات $Y_1 , Y_2 , ... , Y_q$ التي لها أكبر أوزان قويمة b_{ij} بغض النظر عن إشارة هذه الأوزان موجبة كانت أم سالبة.

ومن المفيد هنا أن نذكر بأنه قد نجد الوزن القويم a_{i1} بإشارة موجبة بينما في الوقت نفسه يكون معامل الإرتباط (أو ما سنطلق عليه بالمعاملات التركيبية Structure Coefficients) ما بين U_i والمتغير X_1 سالبة (أي عكس ذلك). وتفسيرنا لظهور هذه الحالة العكسية يعود إلى وجود إرتباط عالي ما بين المتغير X_1 وأي عدد من المتغيرات الأخرى في مجموعة X مما يدفع إلى ظهور هذه الحالة. أي هنالك حالة التعدد الخطي (multicollinearity) ما بين هذه المجموعة من المتغيرات وتأثيرها هنا مشابه لتأثيرها على تقدير المعاملات في نموذج الإنحدار المتعدد. وفي مثل هذه الحالة، لا يمكن الوقوف على الحجم الحقيقي لإسهام هذا المتغير بشكلٍ مستقل تجاه المتغير القويم بل هنالك تداخل من تأثيرات المتغيرات الأخرى التي له معها إرتباطات واضحة. ولتجاوز هذه الحالة، فإننا نرى أهمية النظر إلى معامل الإرتباط (المعاملات التركيبية) ما بين المتغير القويم وكل متغير أولي مقابل له بدلاً من الأوزان القويمة a_{ij} أو b_{ij} . والمعاملات التركيبية $\mathbf{S_{xi}}$ و $\mathbf{S_{yi}}$ يمكن حسابها حسب الصيغة التالية:

$$\mathbf{S_{xi}} = Rxx\,(a_i) = A\,(a_i)$$

$$\mathbf{S_{yi}} = Ryy\,(b_i) = B\,(b_i)$$

وفي ضوء هذه المعادلات وفي حالة كون مصفوفة الإرتباط لأي مجموعة عبارة عن مصفوفة الوحدة (Identity matrix)، فإنه من الواضح أن المعامل التركيبي لأي متغير

أصيل يكون مساوياً للوزن القويم لذلك المتغير. وبشكلٍ عام، فإن مربع المعامل التركيبي لأي متغير يمثل نسبة مساهمته في تفسير التباين الحاصل في المتغير القويم.

والمثال التالي مأخوذ عن كتاب الطرق الإحصائية متعددة المتغيرات لمؤلفه B. F. J. Manly ويتضمن خمسة متغيرات جينية وتمثل مجموعة X وأربع متغيرات بيئية وتمثل مجموعة Y لنوع معين من الفراشات(6). وقد تم جمع البيانات من 16 مستعمرة بولايتين من الولايات الأمريكية (كاليفورنيا وأوريغون). والجدول التالي يبين القيم المعيارية لمجموعتي المتغيرات، ولغرض الترتيب المريح، فإنه يفضل أن تكون مجموعة المتغيرات الأكبر عدداً ممثلة بالمجموعة X والأخرى الأقل عدداً بالمجموعة Y.

المتغيرات الجينية					المتغيرات البيئية			
X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Y_4
-0.42	-0.56	0.32	0.29	-0.02	-0.56	1.07	0.12	-0.35
-0.42	1.18	0.13	-0.66	-0.38	-0.45	-0.58	-0.82	1.02
-0.42	-0.16	0.88	-0.26	-0.02	-0.53	0.00	0.12	0.47
-0.42	-0.43	0.04	-0.21	0.91	-0.54	0.00	0.12	0.47
-0.42	-0.83	-0.98	-0.06	1.65	-0.54	0.00	0.12	0.47
-0.42	-0.69	0.04	-0.36	1.37	-0.60	-0.93	0.27	0.65
-0.42	-0.96	-0.33	-0.06	0.91	-0.41	-0.50	0.27	0.65
1.99	1.84	1.99	-1.31	-1.22	-0.50	-1.29	0.59	0.56
2.96	2.51	1.25	-1.16	-1.59	-0.52	-1.29	0.59	0.56
-0.42	-0.83	-1.17	1.44	-0.48	-0.21	-0.65	0.27	0.19
-0.18	-0.43	1.43	-0.91	0.45	-0.12	-0.43	0.59	0.56
-0.42	-0.03	-0.42	0.74	-0.39	-0.04	2.14	0.43	-0.26
-0.42	0.24	-0.33	-0.21	0.35	0.14	0.42	0.74	-0.44
0.30	-0.03	-0.14	-0.96	0.91	-0.04	-0.50	1.21	-0.08
-0.42	-0.29	-1.07	1.64	-1.22	2.00	1.00	-2.08	-1.45
-0.42	-0.56	-1.63	2.03	-1.22	2.92	1.57	-2.55	-3.00

ومنها يتم تحديد مصفوفة معامل الإرتباط التالية ما بين مجموعتي المتغيرات.

$$
\begin{bmatrix}
1.000 & 0.855 & 0.618 & -0.532 & -0.506 & \vdots & -0.203 & -0.530 & 0.295 & 0.221 \\
0.855 & 1.000 & 0.615 & -0.548 & -0.597 & \vdots & -0.190 & -0.410 & 0.173 & 0.246 \\
0.618 & 0.615 & 1.000 & -0.824 & -0.127 & \vdots & -0.573 & -0.550 & -0.536 & 0.593 \\
-0.532 & -0.548 & -0.824 & 1.000 & -0.264 & \vdots & 0.727 & 0.699 & -0.717 & -0.759 \\
-0.506 & -0.597 & -0.127 & -0.264 & 1.000 & \vdots & -0.458 & -0.138 & 0.438 & 0.412 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \vdots & \cdots & \cdots & \cdots & \cdots \\
-0.203 & -0.190 & -0.573 & 0.727 & -0.458 & \vdots & 1.000 & 0.568 & -0.828 & -0.936 \\
-0.530 & -0.410 & -0.550 & 0.699 & -0.138 & \vdots & 0.568 & 1.000 & -0.479 & -0.705 \\
0.295 & 0.173 & -0.536 & -0.717 & 0.438 & \vdots & -0.828 & -0.479 & 1.000 & 0.719 \\
0.221 & 0.246 & 0.593 & -0.759 & 0.412 & \vdots & -0.936 & -0.705 & 0.719 & 1.000
\end{bmatrix}
$$

وهذا يمثل المصفوفة:

$$\begin{bmatrix} A & \vdots & C \\ \cdots & \cdots & \cdots \\ C' & \vdots & B \end{bmatrix}$$

وبعد تطبيق المعادلة (1) في أعلاه نتوصل أولاً إلى القيم المميزة التالية:

0.7731 , 0.5570 , 0.1694 , 0.0472

ومنها يتم تطبيق المعادلتين (2 ، 3) ليتم تحديد قيم الأوزان القويمة **b** ومن ثم قيم **a** المقابلة لنخرج بالمعادلات الخطية التالية التي تمثل كامل أزواج المتغيرات القويمة بقيم قياسية:

$$U_1 = -0.675X_1 + 0.909X_2 + 0.367X_3 + 1.442X_4 + 0.269X_5$$
$$V_1 = -0.114Y_1 + 0.619Y_2 - 0.693Y_3 + 0.048Y_4$$

$$U_2 = -1.087X_1 + 3.034X_2 + 2.216X_3 + 3.439X_4 + 2.928X_5$$
$$V_2 = -0.777Y_1 + 0.980Y_2 - 0.562Y_3 + 0.928Y_4$$

$$U_3 = 1.530X_1 + 2.049X_2 + 2.231X_3 + 4.916X_4 + 3.611X_5$$
$$V_3 = -3.654Y_1 - 0.601Y_2 - 0.565Y_3 - 0.0483.623Y_4$$

$$U_4 = 0.284X_1 - 2.331X_2 - 0.867X_3 - 1.907X_4 - 1.133X_5$$
$$V_4 = 1.594Y_1 + 0.860Y_2 + 1.599Y_3 + 0.742Y_4$$

ومن خلال القيم المميزة λ_i في أعلاه يتم تحديد الإرتباطات القويمة Rci والتي تساوي الجذر التربيعي للقيم المميزة حيث ستكون:

$$Rc_1 = 0.879 \quad , \quad Rc_2 = 0.746 \quad , \quad Rc_3 = 0.412 \quad , \quad Rc_4 = 0.217$$

ومع أن قيم الإرتباطين الأول والثاني تبدوان كبيرة نسبياً إلا أنه بتطبيق إختبار بارتليت لم تكن أياً منها ذات معنوية بمستوى (0.05). ومع أننا يمكن أن نكتفي بإختبار الأولى لأنها غير معنوية ونتوقف، إلا أننا سنذكر هنا جميع نتائج الإختبار هذا حيث أن

($\Delta_0^2 = 27.85$ بدرجات حرية 20) مقارنة مع 31.4104 = χ^2_{20} الجدولية

($\Delta_1^2 = 11.53$ بدرجات حرية 12) مقارنة مع 21.0261 = χ^2_{12} الجدولية

($\Delta_2^2 = 2.57$ بدرجات حرية 6) مقارنة مع 12.5916 = χ^2_{6} الجدولية

($\Delta_3^2 = 0.52$ بدرجات حرية 2) مقارنة مع 5.99147 = χ^2_{2} الجدولية

وقد يبدو من الغريب أن نحصل على نتيجة غير معنوية مع قيمة الإرتباط القويم الأول الذي هو كبير بشكلٍ واضح. والسبب، وكما ذكرنا سابقاً، هو نتيجة مباشرة لتأثير حجم العينة الصغير نسبياً $(n=16)$. ومع ذلك لعله من المفيد أن نضع عدم المعنوية هذه جانباً لغرض المضي في استكمال تفسير النتائج بالنسبة للإرتباط القويم الأول (Rc_1).

وقبل البدء في ذلك لعله من المفيد أن نذكر هنا قيم معامل الإرتباط (المعامل التركيبية) ما بين الزوج الأول من المتغيرات القويمة (U_1 , V_1) وكل متغير أولي مقابل لهما سوية مع الأوزان القويمة المرافقة في الجدول التالي:

المعامل بالنسبة إلى U_1 مع مجموعة X			المعامل بالنسبة إلى V_1 مع مجموعة Y		
المتغير	الوزن القويم	الإرتباط S_{xi} (المعامل التركيبية) بين U_1 و X	المتغير	الوزن القويم	الإرتباط S_{yi} (المعامل التركيبية) بين V_1 و Y
X_1	−0.675	−0.57	Y_1	−0.114	0.77
X_2	0.909	−0.39	Y_2	0.619	0.85
X_3	0.367	−0.70	Y_3	−0.693	−0.86
X_4	1.442	0.92	Y_4	0.048	−0.78
X_5	0.269	−0.36			

ومن خلال هذا الجدول وبالنسبة إلى U_1 نلاحظ أن الوزن القويم بالنسبة إلى X_1 هو الوحيد بإشارة سالبة مما نستطيع القول بأن إتجاه قيم X_1 معاكسة تماماً لإتجاه قيم المتغيرات الأخرى في المجموعة. ومن جهة أخرى، وبالنسبة إلى V_1 فإننا نلاحظ أوزان قويمة عالية موجبة مع Y_2 (0.619) وعالية سالبة مع Y_3 (-0.693) مما يمكننا قوله بأن القصور في قيم المتغير الجيني X_1 في مستعمرة ما للفراشات يتزامن مع تزايد في قيم المتغير البيئي Y_2 وتناقص قيم المتغير البيئي Y_3.

وبالنظر إلى معامل الإرتباط ما بين المتغير القويم U_1 ومجموعة المتغيرات **X** نلاحظ أن U_1 يرتبط إيجابياً بشكلٍ واضح مع المتغير X_4 (0.92) وسلبياً مع بقية المتغيرات الأربعة الأخرى والتي أعلاها مع X_3 (-0.70)**.** وبناءً على ذلك يمكننا القول بأن U_1 يؤشر إرتفاع قيمة المتغير X_4 . وهذا بطبيعة الحال يعتبر مغايراً بعض الشئ عما يمكننا قوله من خلال النظر إلى الأوزان القويمة لمجموعة المتغيرات **X** ضمن الدالة U_1. وبشكلٍ عام، فإن التفسير في ضوء معامل الإرتباط (المعامل التركيبية) يبدو أفضل هنا مما هو عليه مع الأوزان القويمة.

وجدير بالذكر هنا أن هنالك مشكلة حقيقية في تفسير النتائج بالنسبة إلى المتغيرات القويمة في حالة وجود إرتباط عالي نسبياً ما بين المتغيرات الأصلية وهذا ما وجدناه هنا بالفعل حيث

معامل الإرتباط العالي نسبياً ما بين المتغيرات في المجموعة **X** وكما يتضح لنا من الجدول الأخير.

مقاييس أخرى للتحليل

هنالك مقياسان آخران يمكن حسابهما من البيانات وقد يضيفان بعض الجوانب الأخرى من تفسير النتائج وهما:

1- معامل كفاية الجودة (A_d) (Adequacy Coefficint)

وهذا المعامل يشير إلى نسبة التباين الكلي الحاصل لمجموعة المتغيرات الأصلية والمفسرة من قبل المتغير القويم لتلك المجموعة. ويتم حساب هذه المعامل والتي تساوي معدل مجموع مربعات المعامل التركيبية لمجموعة المتغيرات الأصلية عند متغير قويم معين ووفقاً للصيغة التالية:

$$A_d\left(U_{xi}\right)=\frac{\sum_{r=1}^{p} S_{xir}^2}{p}(100)$$

$$A_d\left(V_{yi}\right)=\frac{\sum_{r=1}^{q} S_{yir}^2}{q}(100)$$

ولأننا إعتمدنا معامل الإرتباط القويم الأول، فإن معامل كفاية الجودة لمجموعتي المتغيرات هما حسب الآتي:

$$A_d\left(U_{x1}\right)=\frac{\sum_{r=1}^{5} S_{x1r}^2}{5}(100)=\frac{(-0.57)^2+(-0.39)^2+(-0.70)^2+(0.92)^2+(-0.36)^2}{5}=0.40$$

$$A_d\left(V_{y1}\right)=\frac{\sum_{r=1}^{4} S_{y1r}^2}{4}(100)=\frac{(0.77)^2+(0.85)^2+(-0.86)^2+(-0.78)^2}{4}=0.67$$

وهذا يعني أن المتغير القويم U يوضح ما نسبته 40% من التباين الكلي لمجموعة المتغيرات الجينية **X** . ووفقاً لقيم S_{x1r}^2 فإنه بإستطاعتنا ترتيب المتغيرات الجينية **X** حسب قوة تأثيرها في المتغيرات البيئية **Y** فيكون المتغير X_4 هو الأكثر تأثيراً (المرتبة الأولى) يليه المتغير X_3 ثم X_1 ثم X_2 ثم الأقل تأثيراً X_5 . كما أن المتغير القويم V يوضح ما نسبته 67% من التباين الكلي لمجموعة المتغيرات البيئية **Y** لهذه الفراشات.

2- **معامل الإفاضة** $Rd_{y|x}$ **(Redundancy Coefficiont)**

وهذا المعامل يمثل نسبة التباين الحاصلة في متغيرات مجموعة معينة والمفسرة من قبل متغيرات المجموعة الأخرى . أي أنه يفيد في معرفة مدى تأثير مجموعة متغيرات أصلية **X** في مجموعة المتغيرات الأخرى **Y** (والعكس بالعكس). وقيمة هذا المعامل تساوي معدل مجموع مربعات معاملات الإرتباط المتعدد وقيمته ضمن المجال [1 , 0] ويأخذ القيمة 1.0 لأي زوج من المتغيرات القويمة عندما تكون نسبة التباين المشترك لها مساوياً إلى (100%) أي عندما يكون الإرتباط القويم عند هذا الزوج مساويا إلى(1.0). وقد تحسب هذه القيمة أيضاً كما في الصيغة التالية:

$$Rd_{y|x} = (Rc_1)^2 . \ A_d(V_{y1}) = (0.879)^2 . (0.67) = 0.52$$

وهذا يعني أن مجموعة المتغيرات الجينية **X** توضح ما نسبته 52% من تباين مجموعة المتغيرات البيئية **Y** للفراشات.

ملاحظة:

نود التذكير هنا بأن الحسابات أعلاه مبنية على أساس إطلاق المجموعة **X** للمجموعة ذات العدد الأكبر من المتغيرات وعددها (p) بالنسبة للمجموعتين و (q<p).

وفي هذا السياق، نود أن نذكر هنا نتائج مثال تطبيقي آخر لإستخدام تحليل الإرتباط القويم من خلال دراسة أدت نتائجها في حينه إلى قيام وزارة التعليم العالي العراقية لإعادة النظر في بعض أسس القبول في كليات المجموعة الطبية[7]. وخلاصة فكرة الدراسة، التي أجريت عام 1998، أن وزارة التعليم العالي قد حددت شروط خاصة للقبول في كليات المجموعة الطبية (كلية الطب، كلية طب الأسنان، كلية الصيدلة وكلية الطب البيطري) بتحديد مواد المفاضلة وهي (الفيزياء، الكيمياء، الأحياء) وأن لا تقل درجة الطالب في أي منها عن 70% في المرحلة الثانوية (التوجيهي)، والتي تم تطبيقها منذ العام الدراسي 1972/1971، منطلقين من الإفتراض بأن هذه المواد الثلاثة لها تأثير على تحصيل الطالب في مواد السنة الأولى في هذه الكليات وبالتالي في مسيرته للسنوات الأخرى. ولذلك تم تصميم الدراسة موضوع البحث لغرض التأكد من منطقية هذا الإفتراض من عدمه ومن خلال نتائج الإرتباط القويم بين درجات الطالب في مواد المرحلة الثانوية (مجموعة المتغيرات X) ودرجاته في مواد السنة الأولى من الكلية المنتسب اليها (مجموعة المتغيرات Y) ولثلاث سنوات أكاديمية (1986/85، 1990/89، 1994/93). وسنكتفي بتناول بيانات كلية الطب فقط لغرض التوضيح فيما تشيراليه النتائج والتي تظهر في الجدول التالي:

المتغير	مواد المرحلة الثانوية	السنة الدراسية					
		1986/1985		1990/1989		1994/1993	
		قيمة % $(Sxi)^2$	ترتيب العلاقة	قيمة % $(Sxi)^2$	ترتيب العلاقة	قيمة % $(Sxi)^2$	ترتيب العلاقة
X_1	لغة عربية	14	6	50	2	34	6
X_2	لغة إنكليزية	24	5	24	5	49	3
X_3	رياضيات	52	3	52	1	61	2
X_4	أحياء	56	2	46	3	38	5
X_5	كيمياء	50	4	37	4	62	1
X_6	فيزياء	59	1	19	6	46	4

وفي ضوء النتائج هذه، فإن مدى تحقق صحة إفتراض الوزارة يتحقق من خلال ترتيب علاقة متقدم لمواد المفاضلة (فيزياء، كيمياء، أحياء) في جميع هذه السنوات. أي أن تأخذ الترتيب (1، 2، 3). ولكن الذي حصل أن هذه المواد أخذت الترتيبات (1، 4، 2) و (6 ، 4، 3) و (4، 1، 5) للسنوات الثلاث على التوالي. أي أن مادة الرياضيات أخذت الترتيب المتقدم بدلاً من الكيمياء لعام 1986/85، ومادتي الرياضيات واللغة العربية بدلاً من الفيزياء والكيمياء لعام 1990/89، ومادتي الرياضيات واللغة الإنكليزية بدلاً من الفيزياء والأحياء لعام 1994/93. ووفقاً لذلك، يصبح من المؤكد عدم صحة الإفتراض الذي إعتمدته الوزارة بشأن منطقية مواد المفاضلة.

التحليل العاملي

Factor Analysis (FA)

التحليل العاملي عبارة عن إسلوب غالباً ما يستخدم في تكوين متغيرات جديدة تلخص جميع المعلومات التي من الممكن توفرها في المتغيرات الأصلية. وعلى سبيل المثال، في حالة إعطاء إمتحان لطلبة مرحلة تعليمية محددة في عدد من المواد مثل القراءة والإملاء والرياضيات والعلوم في الوقت الذي يحصل فيه الطلبة في النهاية على تقدير عالي، متوسط، أو واطئ كنتيجة نهائية لجميع المواد. فإذا ما كان هذا الذي سيحصل، فإننا نستطيع القول بأن نتائج الإمتحان هذه يمكن توضيحها من خلال تحديد سمة أو عامل مشترك لجميع نتائج الإمتحانات الأربعة. في هذا المثال، قد يبدو منطقياً لإفتراض مثل هذه السمة (العامل) بمثابة تعبير عن الذكاء أو الأداء العام.

والتحليل العاملي يستخدم أيضاً لدراسة العلاقات التي من الممكن وجودها ما بين المتغيرات التي تم قياسها ضمن مجموعة البيانات. ومثلما هي الحال في تحليل المركبات الرئيسية، فإن التحليل العاملي هو من أساليب تحكم المتغيرات.

أحد الأهداف الرئيسية للتحليل العاملي يكمن في تحديد ما إذا كانت المتغيرات الناتجة تعكس أنماطاً من العلاقات مع بعضها البعض بحيث يمكن تقسيم المتغيرات إلى مجاميع جزئية من المتغيرات تكون مترابطة بشكلٍ واضح فيما بينها ضمن مجموعة بعينها في حين نجد هذه المتغيرات أقل ترابطاً ضمن المجاميع الأخرى. ولذلك فإن التحليل العامليغالباً ما يستخدم لدراسة البناء الترابطي ما بين المتغيرات ضمن مجموعة البيانات. وهذه المتغيرات كمية وبالتالي فإنها ذات مقياس نسبي أو فئوي.

أحد أوجه الشبه ما بين تحليل المركبات الرئيسية والتحليل العاملي هو أن التحليل العاملي يمكن استخدامه أيضاً لتكوين متغيرات جديدة غير مترابطة مع بعضها البعض. مثل هذه المتغيرات تسمى الدرجات العاملية Factor Scores.

يتميز التحليل العاملي بإحدى الإيجابيات مقارنة بتحليل المركبات الرئيسية عند تكوين المتغيرات الجديدة وهي أن هذه المتغيرات الجديدة التي تتكون من خلال التحليل العاملي هي، بشكلٍ عام، قابلة للتفسير بشكلٍ أسهل بكثيرمقارنة بتلك المكونة من خلال تحليل المركبات الرئيسية. وهذا يعني أنه إذا ما أراد الباحث تكوين مجموعة متغيرات جديدة أصغر عدداً وقابلة للتفسير والإستنتاج وتلخص معظم المعلومات في المتغيرات الأصلية، فإنه يجب أخذ التحليل العاملي في الإعتبار حيث عدد العوامل الناتجة تكون أقل عدداً من المتغيرات التي نبدأ التحليل بها.

أهداف التحليل العاملي

1) التعرف على الأنماط البينية

فإذا كانت لدينا مصفوفة إرتباطات بين مجموعة من المتغيرات تمثل خصائص معينة فإن اسلوب التحليل العاملي يكشف عن الأنماط المنفصلة للعلاقات البينية التي تتضمنها المتغيرات ويحدد علاقة كل متغير بتلك الأنماط ودرجة هذه العلاقة.

2) الإختصار في وصف البيانات

إذا كان لدينا مجموعة كبيرة من المشاهدات التي تخص عدد كبير من المتغيرات، فإنه يمكن أن يتم تركيز هذه البيانات ضمن عدد قليل من العوامل تقوم مقام المتغيرات العديدة في إجراء الوصف والمقارنة.

3) بناء مقاييس التقدير

كثيراً ما يتطلب الأمر تصميم مقياس لتقدير سلوك الأفراد في مجال معين، ويستلزم ذلك إعطاء أوزان معينة للخصائص التي يتضمنها هذا المقياس. والتحليل العاملي يحقق هذا الهدف بتصنيفه لهذه الخصائص والمتغيرات في صورة عوامل مستقلة.

4) تحويل البيانات

يستعمل التحليل العاملي في تحويل البيانات إلى صورة أخرى تتوفر فيها بعض الشروط التي يمكن من خلالها تطبيق أساليب إحصائية أكثر جدوى على هذه البيانات. ومثال ذلك في تحليل الإنحدار المتعدد فإن التقدير الصحيح للمعاملات يتطلب عدم وجود حالة التعدد الخطي. ولكن في حالة وجود مثل هذه الحالة، فإنه يمكن إستخدام التحليل العاملي وتحويلها إلى أقل عدداً من العوامل غير المترابطة يمكن إحلالها محل المتغيرات الأصلية في معادلة الإنحدار.

النموذج العاملي Factor Model

إذا افترضنا نظام متعدد المتغيرات يتضمن عدد P من الإستجابات (المتغيرات العشوائية)

$[X_1, X_2, ..., X_p]$ تتبع توزيعاً طبيعياً متعدداً، فإنه يمكن أن نكتب النموذج التالي للإستجابات:

$$X_1 = a_{11}F_1 + a_{12}F_2 + ... + a_{1m}F_m + e_1$$
$$\vdots$$
$$X_p = a_{p1}F_1 + a_{p2}F_2 + ... + a_{pm}F_m + e_p$$

أو بشكلٍ عام:

$$X_i = a_{i1}F_1 + a_{i2}F_2 + ... + a_{im}F_m + e_i \quad ....(1)$$

وبصيغة المصفوفات تكون:

$$\underline{X} = A\underline{F} + \underline{e} \qquad .... (2)$$

$$\underline{X'} = [X_1, X_2, ..., X_p]$$
$$\underline{F'} = [F_1, F_2, ..., F_p]$$
$$\underline{e'} = [e_1, e_2, ..., e_p]$$
$$A = \begin{bmatrix} a_{11} & \cdots\cdots & a_{1m} \\ \vdots & \cdots . \cdots & \\ a_{p1} & \cdots\cdots & a_{pm} \end{bmatrix}$$

حيث أن:

$\underline{X}$: موجه المتغيرات العشوائية

$\underline{F}$: موجه العوامل المشتركة Common Factors

A : مصفوفة تحميلات العوامل Factor Loadings

$\underline{e}$: الموجه العشوائي

الفروض الأساسية للتحليل العاملي

1- <u>الفرضية الأولى</u>

تعتمد هذه الفرضية على أساس وجود إرتباطات بين المتغيرات قيد الدراسة نتيجة وجود عوامل مشتركة فيما بينها، ويكون النموذج العاملي إلى (p) من المتغيرات المشاهدة لعينة حجمها (n) على أساس تكوين دالة خطية إلى (q) من العوامل (بالقيمة المعيارية) وكما يلي:

$$Z_{ij} = a_{1j}F_1. + a_{2j}F_2 + ... + a_{qj}F_q + d_j u_{ij} \qquad .... (3)$$

حيث أن:

Z_{ij} = القيمة المعيارية للمشاهدة (i) بالنسبة للمتغير (j)

a_{qj} = تحميل العامل (بالنسبة للمتغير .(j) وهي أوزان مرافقة لقيم العوامل المشتركة)

F_q = القيمة المعيارية للعامل (q) المشترك المحدد

u_{ij} = القيمة المعيارية للمفردة (i) للعامل المشترك المحدد

d_j = معامل يمثل الوزن المرافق لقيمة العامل الفريد (العامل الخاص بمتغيرواحد).

ويقسم التباين الكلي للمتغيرات إلى ثلاثة أنواع (ضمن هذه الفرضية) وهي:

1. التباين المشترك *Common Variance* : وهو الجزء الذي يرتبط مع بقية المتغيرات الأخرى من خلال العوامل المشتركة.
2. التباين الخاص *Specific Variance* : وهو الجزء الذي لا يرتبط مع بقية المتغيرات بل مع المتغير نفسه.
3. تباين الخطأ *Error Variance* : وهو الجزء الناتج من خلال حدوث أخطاء عند سحب العينة أو قياسها أو تغيرات أخرى ترتبط بالشخص الذي يسحب المشاهدة.

ويمكن التعبير عن التباين الكلي σ_j^2 للمتغير وأجزائه كما يلي:

$$\sigma_j^2 = (\sigma_{j1}^2 + \sigma_{j2}^2 + ...+ \sigma_{jq}^2) + (\sigma_{js}^2) + (\sigma_{je}^2)$$
$$= (Common\ Variance) = (SpecificVariance) + (ErrorVariance) \quad .... (4)$$

ويساهم كلاً من التباين المشترك والتباين الخاص في تكوين التباين الثابت (المعتمد) كما يفترض بأن تباين الخطأ لا يرتبط بالتباين الثابت.

وبقسمة طرفي المعادلة (4) على σ_j^2 يصبح لدينا:

$$\frac{\sigma_j^2}{\sigma_j^2} = \frac{(\sigma_{j1}^2 + \sigma_{j2}^2 + ...+ \sigma_{jq}^2}{\sigma_j^2} + \frac{\sigma_{js}^2}{\sigma_j^2} + \frac{\sigma_{je}^2}{\sigma_j^2} \quad .... (5)$$

$$1 = a_{j1}^2 + a_{j2}^2 + ...+ a_{jq}^2 + S_j^2 + e_j^2 \quad .... (6)$$

حيث أن الجذر التربيعي للتباينات المشتركة $a_{j1}^2 , a_{j2}^2 , ... , a_{jq}^2$ هي تحميلات العوامل والتي تمثل مقدار الإرتباط للمتغير (j) بكل عامل.

2- <u>الفرضية الثانية</u>

تقوم هذه الفرضية على أساس أن معامل الإرتباط بين متغيرين (i , j) يعود إلى طبيعة تشبعهما بالعوامل المشتركة ومدى هذا التشبع. أي أن معامل الإرتباط بين متغيرين يساوي مجموع حاصل ضرب تحميلات المتغيرات بالعوامل المشتركة بينهما. أي أن

$$r_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + ... + a_{iq}a_{jq} \quad .... (7)$$

حيث أن هذه المعادلة إنما تعبر عن العوامل المتعامدة (orthogonal) ويمكن كتابتها بصيغة المصفوفات وبالشكل التالي:

$$R = AA'$$

حيث أن:

R = مصفوفة الإرتباط

A = مصفوفة تحميلات العوامل

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \cdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{bmatrix}$$

الإشتراكيات (Communalities) وطرق تقديرها

إن قيم الإشتراكيات (الشيوع) يرمز لها بالرمز h_j^2 حيث $0 \le h_j^2 \le 1$ والتي تكون قيمتها

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots\ldots\ldots\ldots + a_{jq}^2 \quad , \quad j = 1 , 2 , \ldots\ldots, q$$

أي أنها تمثل مجموع مربعات تحميلات (تشبعات) كل متغير وهو بمثابة نسبة التباين الذي تفسره العوامل المشتركة الناتجة من تحليل مصفوفة الإرتباط R .

وفي ضوء ذلك، يمكننا وضع الصيغة لتالية للمعادلة (6)

$$1 = h_j^2 + S_j^2 + e_j^2 = R_{jj} + e_j^2$$

وهنا سنعمل على تقدير قيم الإشتراكيات لتحل كعناصر تقديرية في مصفوفة الإرتباط. فإذا كانت العناصر القطرية لمصفوفة الإرتباط تساوي الواحد صحيح فإننا نطلق عليها تسمية "مصفوفة الإرتباطات الكاملة" (Complete correlation matrix). أما إذا تم استبدال العناصر القطرية بقيم الإشتراكيات h_j^2 فإنها تسمى "مصفوفة الإرتباطات المخفضة" (Reduced correlation matrix)

طرق تقدير الإشتراكيات

1) الإرتباط الأكبر (maximum correlation)

وفي هذه الطريقة يتم اعتماد أكبر معاملات الإرتباط بين متغيرما موضوع التقدير وبقية المتغيرات الأخرى لتمثل القيم التقريبية للإشتراكيات ويفضل إستخدام هذه الطريقة في حالة مصفوفة ارتباطات مكونة من عدد كبير من المتغيرات.

2) ثلاثي الأبعاد (third dimension)

ووفقاً لهذه الطريقة يتم تقدير اشتراكية أي متغير (j) بالصيغة الثلاثية:

$$h_j^2 = \frac{r_{ji} r_{jk}}{r_{ik}}$$

حيث أن (i) و (k) هما المتغيران اللذان لهما أعلى إرتباط مع المتغير (J). ووفقاً لهذه الطريقة يتم تقليل تأثير الإرتباطات العالية.

3) **معدل الإرتباطات (mean of correlations)**

ويتم من خلال استخراج معدل معامل الإرتباط لذلك المتغير (j) مع بقية المتغيرات وعلى النحو التالي:

$$h_j^2 = \sum_{j \neq i} \frac{r_{ji}}{n-1}$$

4) مربع الإرتباط المتعدد Square Multiple Correlation (SMC)

وتعتبر هذه الطريقة من أكثر الطرق استخداماً لتقدير قيم الإشتراكيات h_j^2 وبالإعتماد على مصفوفة الإرتباط R حيث يتم تحديد معكوس هذه المصفوفة R^{-1} واستخدامها لتقدير الإشتراكيات على النحو التالي:

$$h_j^2 = SMC = 1 - \frac{1}{r_{jj}}$$

حيث أن r_{jj} تمثل العنصر القطري لمعكوس مصفوفة الإرتباطات لذلك المتغير (j).

ولغرض توضيح إستخدام البعض من هذه الطرق، سنعتمد مثالاً بثلاثة متغيرات لغرض التبسيط.

مثال:

لنفترض أن لدينا مصفوفة الإرتباطات التالية للمتغيرات X_1 , X_2 , X_3 :

$$R = \begin{bmatrix} 1 & 0.61 & 0.40 \\ 0.61 & 1 & 0.48 \\ 0.40 & 0.48 & 1 \end{bmatrix}$$

فإن تقدير الإشتراكيات الثلاثة ستكون وفقاً للطريقة الأولى كما يلي:

$$h_1^2 = 0.61$$
$$h_2^2 = 0.61$$
$$h_3^2 = 0.48$$

ووفقاً للطريقة الثانية فإنها:

$$h_1^2 = \frac{(0.61)(0.40)}{0.48} = 0.508$$
$$h_2^2 = \frac{(0.61)(0.48)}{0.40} = 0.732$$
$$h_3^2 = \frac{(0.40)(0.48)}{(0.61)} = 0.314$$

وبالنسبة للطريقة الثالثة:

$$h_1^2 = \frac{(0.61)+(0.40)}{2} = 0.550$$
$$h_2^2 = \frac{(0.61)+(0.48)}{2} = 0.545$$
$$h_3^2 = \frac{(0.40)+(0.48)}{2} = 0.440$$

وبالنسبة للطريقة الرابعة نحتاج أولاً إلى
تحديد معكوس المصفوفة وحسب الآتي:

$$adjR = \begin{bmatrix} 0.77 & -0.42 & -0.11 \\ -0.42 & 0.84 & -0.24 \\ -0.11 & -0.24 & 0.63 \end{bmatrix}$$

$$|R| = 0.47$$

$$R^{-1} = \begin{bmatrix} 0.77 & -0.42 & -0.11 \\ -0.42 & 0.84 & -0.24 \\ -0.11 & -0.24 & 0.63 \end{bmatrix} / 0.47 = \begin{bmatrix} 1.64 & -0.89 & -0.23 \\ -0.89 & 1.78 & -0.51 \\ -0.23 & -0.51 & 1.34 \end{bmatrix}$$

وبالتالي يكون لدينا:

$$h_1^2 = 1 - \frac{1}{r_{11}} = 1 - \frac{1}{1.64} = 0.390$$
$$h_2^2 = 1 - \frac{1}{r_{22}} = 1 - \frac{1}{1.78} = 0.438$$
$$h_3^2 = 1 - \frac{1}{r_{33}} = 1 - \frac{1}{1.34} = 0.254$$

ونلاحظ أن هنالك تبايناً واضحاً في قيم الإشتراكيات المقدرة بهذه الطرق الأربع وهذا بطبيعة الحال بسبب كون عدد المتغيرات صغير ونتوقع أن لا نجد مثل هذا الحجم من الفروقات عند التعامل مع عدد كبير نسبياً من المتغيرات.

حساب مصفوفة الإرتباط

بطبيعة الحال، نحن بحاجة إلى وجود مصفوفة الإرتباط ويمكن تحديد هذه المصفوفة بالشكل الإعتيادي أو من خلال تحميلات العوامل والتي سنوضحها فيما يلي:

لنفترض أنه لدينا حالة أربعة متغيرات وهي X_1 , X_2 , X_3 , X_4 وحصلنا على عاملين معنويين F_1 , F_2 بالتحميلات التالية من قيم معيارية:

$$F=\begin{bmatrix} 0.66 & 0.27 \\ 0.79 & 0.33 \\ 0.34 & 0.64 \\ 0.17 & 0.78 \end{bmatrix} \quad (\text{columns } F_1,\ F_2)$$

$$F*F' = \begin{bmatrix} 0.50 & 0.61 & 0.40 & 0.32 \\ 0.61 & 0.73 & 0.48 & 0.39 \\ 0.40 & 0.48 & 0.52 & 0.55 \\ 0.32 & 0.39 & 0.55 & 0.63 \end{bmatrix}$$

وهذه المصفوفة عبارة عن مصفوفة الإرتباط للمتغيرات الأربعة مع احلال الإشتراكيات لمواقع العناصرالقطرية في المصفوفة. ولكي نثبت ذلك دعنا نحسب قيم الإشتراكيات الأربعة بالشكل الإعتيادي كونها تساوي مجموع مربعات تحميلات العوامل وهي:

$$h_1^2 = (0.66)^2 + (0.27)^2 = 0.50$$

$$h_2^2 = (0.79)^2 + (0.33)^2 = 0.73$$

$$h_3^2 = (0.34)^2 + (0.64)^2 = 0.52$$

$$h_4^2 = (0.17)^2 + (0.78)^2 = 0.63$$

حيث أن الإشتراكية لمتغير ما عبارة عن نسبة التباين لذلك المتغير(بإعتباره متغيراً معتمداً) التي تم توضيحها من خلال العوامل المشتركة F_1 و F_2 (بإعتبارها متغيرات مستقلة ومتعامدة). وبذلك يمكننا كتابة المعادلات الخطية التالية بمثابة معادلات إنحدار خطية:

$$X_1 = a_{11}F_1 + a_{12}F_2 + d_1U_1 \quad = 0.66F_1 + 0.27F_2 + 0.70U_1$$
$$X_2 = a_{21}F_1 + a_{22}F_2 + d_2U_2 \quad = 0.79F_1 + 0.33F_2 + 0.52U_2$$
$$X_3 = a_{31}F_1 + a_{32}F_2 + d_3U_3 \quad = 0.34F_1 + 0.64F_2 + 0.69U_3$$
$$X_4 = a_{41}F_1 + a_{42}F_2 + d_4U_4 \quad = 0.17F_1 + 0.78F_2 + 0.60U_4$$

حيث أن $d_i = \sqrt{1-h_i^2}$

ومن المعلوم أن مربع معامل الإنحدار الجزئي بالوحدات المعيارية هوعبارة عن التباين الموضح للمتغير التابع من خلال المتغير المستقل. وبذلك فإنه من المعادلة الخطية الأولى في أعلاه نجد أن $(0.66)^2 = 43.66\%$ وهذا يعني أن حوالي 44% من تباين المتغير X_1 تم تفسيره من خلال العامل الأول F_1 وأن $(0.27)^2 = 7.29\%$ يعني أن حوالي 7% من تباين المتغير X_1 تم تفسيره من خلال العامل الثاني F_2. وأن مجموعهما حوالي = 7.29 + 43.66 51% وهي نسبة التباين الكلية التي تم تفسيرها من خلال العاملين F_1 و F_2.

كما نلاحظ، ومن خلال تحميلات العوامل، أن العامل الأول هو المحدد المهم بالنسبة للمتغيرين X_1 و X_2 كما أن العامل الثاني هو المحدد المهم بالنسبة للمتغيرين X_3 و X_4.

والمثال التالي[8] يوضح نتائج التحليل العاملي لبيانات دراسة تربوية تتضمن 9 متغيرات كمية تمثل إختبارات تربوية ونفسية واقتصر الإختيار على العوامل الثلاثة الأولى لمعنويتها حيث كانت النتائج لتحميلات هذه العوامل كما في الجدول التالي:

المتغيرات		تحميلات العوامل		
رمزها	ما تمثله	F_3	F_2	F_1
X_1	المتناسبات اللفظية	0.13	0.16	0.83
X_2	سلاسل الأرقام	- 0.04	0.89	0.02
X_3	ذاكرة الأشكال	0.71	0.43	0.01
X_4	فهم الأمثال	0.01	0.21	0.73
X_5	الإستدلال اللفظي	0.15	0.11	0.79
X_6	سلاسل الأشكال	0.68	0.53	0.02
X_7	ذاكرة الأرقام	0.12	0.84	0.02
X_8	إدراك الأشكال	0.84	0.38	0.07
X_9	عمليات حسابية	- 0.01	0.91	0.03

ومن خلال النتائج أعلاه نلاحظ التحميلات العالية لهذه العوامل وما تقابله من متغيرات ضمن كل عامل من هذه العوامل الثلاثة ليستقر الرأي على إعطاء سمة لكل عامل وفقاً لما يتناسب والمتغيرات التي تقابل أكبر التحميلات فيه لتكون كما يلي:

- العامل الأول (F_1) والذي أطلق عليه الباحث عامل (الإدراك اللفظي) لتميزه في:

المتناسبات اللفظية 0.83

الإستدلال اللفظي 0.79

فهم الأمثال 0.73

- العامل الثاني (F_2) والذي أطلق عليه الباحث عامل (الإدراك الذهني) لتميزه في:

عمليات حسابية 0.91

سلاسل الأرقام 0.89

ذاكرة الأرقام 0.84

- العامل الثالث (F_3) والذي أطلق عليه الباحث عامل (الإدراك الشكلي) لتميزه في:

إدراك الأشكال 0.84

ذاكرة الأشكال 0.71

سلاسل الأشكال 0.68

المصادر

1) Johnson, Dallas E. (1998) "Applied Multivariate Methods for Data Analysts" ; Duxbury Press.
2) Boston University, School of Public Health (2013) "Multiple Linear Regression Analysis – seventh examination of the Framingham Offspring Study".
3) Rencher, Alvin C. (2002) "Methods of Multivariate Analysis " 2nd Ed. ; Wiley Interscience.
4) Al-Nsour, Mohannad & Arbaji, Ali (2014) " Obesity and Related Factors Among Jordanian Women of Reproductive Age (Based on Three DHS Surveys, 2002-2012) "; Middle East Health Observatory for Research and Studies (MEHORS).
5) Zaiontz, Charles (2017) "Real Statistics Using Excel"; www.real-statistics.com.
6) Manly, Bryan F. J. (2004) " Multivariate Statistical Methods ; A Primer " ; 3rd Ed. ; Chapman & Hall/CRC.
7) الكبيسي، ماثل كامل (1998) " إستخدام الإرتباط القويم في دراسة العلاقة بين درجات مواد المفاضلة في القبول ودرجات المواد العلمية للسنة الأولى في كليات المجموعة الطبية" /رسالة ماجستير في الإحصاء/ إشراف أ.د. زياد الراوي/الجامعة المستنصرية.
8) العلاق، مهدي محسن اسماعيل (1982) "استخدام التحليل العاملي(طريقة الإمكان الأعظم) في تحليل وتفسير بعض نتائج المسح الجيولوجي في العراق/ رسالة ماجستير في الإحصاء/جامعة بغداد.